EUROPEAN LANGUAGE EQUALITY

D1.18

Report on the Hungarian Language

Authors	Kinga Jelencsik-Mátyus, Enikő Héja, Zsófia Varga, Tamás Váradi
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	D1.18
Deliverable title	Report on the Hungarian Language
Type	Report
Number of pages	26
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
Authors	Kinga Jelencsik-Mátyus, Enikő Héja, Zsófia Varga, Tamás Váradi
Reviewers	Tea Vojtěchová, Sabine Kirchmeier
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

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
ASR	Automatic Speech Recognition
CCTD	Country-Code Top-Level Domain
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DLE	Digital Language Equality
DNN	Deep Neural Network
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
EU	European Union
GPU	Graphics Processing Unit
HPC	High-Performance Computing
LM	Language Model
LT	Language Technology/Technologies
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
PoS	Part-of-Speech
NLG	Natural Language Generation
NLP	Natural Language Processing
NMT	Neural Machine Translation
R&D	Research and Development
SMT	Statistical Machine Translation
TTS	Text-to-speech

Abstract

In the framework of the European Language Equality (ELE) project, the present paper gives a qualitative overview of the current situation of Hungarian Natural Language Processing (NLP). The project's main objectives are to provide a comprehensive landscape of the Hungarian NLP scene by compiling a roadmap of existing language technology tools and datasets in Hungarian, to identify the major gaps in present day national language technologies in the EU as of September 2021, and to determine the essential directions in research and technology. This is part of a joint pan-European effort that will impact the field of language technology (LT) in Europe for the next 10-15 years, including prospective funding.

The large-scale language technology data collection process has aimed at cataloguing, to the largest possible extent, all corpora, lexical and conceptual resources, tools, grammars, and language models available for the Hungarian language as of September 2021. The data collection process was carried out by the Hungarian Research Centre for Linguistics as part of the European Language Equality (ELE) project. Altogether we collected 344 datasets and 180 tools and language models. We hope that our results may be of use for the Hungarian NLP community. The detailed database of the 500+ language technology resources we identified are available for stakeholders online on the ELG website. This work, together with other ELE partner institutions covering over 30 languages in European countries, serves as the basis for a comprehensive proposal and a roadmap for achieving digital language equality in Europe by 2030.

So far, there has been only one study of a similar scope for Hungarian LT. In 2012, the META-NET network and its partner institutions compiled a comprehensive survey of their languages in terms of their LT support and published their findings in a series of White Papers. The present paper is a summary of this new survey that can be considered an update of the book *The Hungarian Language in the Digital Age* (Simon et al., 2012) that was published in the META-NET White Papers series.

In almost a decade that has passed since the publication of Simon and colleagues' work, LT as a field has undergone revolutionary innovations as statistical methods have been abandoned in favour of neural networks. As a result, LT has found its way into our everyday life – we wish to capture these changes as well.

Összefoglaló

A European Language Equality (ELE) projekt egy olyan nagyszabású kezdeményezés, melynek központjában egy stratégiai cselekvési terv kidolgozása áll a digitális nyelvi egyenlőség elérésére Európában 2030-ra. Ehhez a projekt jelenlegi fázisában a részt vevő partnerek feltérképezik, hogy milyen erősségei és milyen hiányosságai vannak nyelveiknek a nyelvtechnológia terén. A projektnek szerves részét képezi egy nagyszabású adatgyűjtés, amely az összes Európai Uniós nyelvre és számos kisebbségi nyelvre kiterjed, és amelynek célja, hogy összegyűjtse az összes, adott nyelvre elérhető természetesnyelv-feldolgozással (NLP) kapcsolatos erőforrást. Így a projekt eredményei közvetetten meghatározzák a nyelvtechnológia európai és magyarországi jövőjét a következő 10-15 évben, ideértve az elérhető Európai Uniós finanszírozási lehetőségeket is. Jelen beszámoló az ELE projekt keretében összeállított, magyar nyelvre kidolgozott nyelvtechnológiai erőforrásokat felölelő adatbázis kvalitatív áttekintése, mely a Nyelvtudományi Kutatóközpont koordinálása alatt jött létre.

Összegyűjtöttük a magyar nyelvre 2021 végén elérhető, saját internetes céloldallal rendelkező nyelvtechnológiai erőforrásokat: korpuszokat, lexikai erőforrásokat, nyelvtechnológiai eszközöket, nyelvtanokat és nyelvmodelleket. Összesen több mint 500 elemből álló adatbázist hoztunk létre, amely 344 adatkészletet és 180 eszközt és nyelvmodellt tartalmaz. A

helyzetfelméréssel együtt egyúttal azonosítottuk a jelenlegi magyar nyelvtechnológiák hiányosságait is, amely meghatározhatja a magyar NLP jövőjének néhány fő kutatási-fejlesztési irányát. Fontos kiemelni, hogy mindez csupán egy pillanatkép egy dinamikusan fejlődő területről. A részletes, többek közt licencekre is kiterjedő adatbázis bárki számára online elérhető az ELG weboldalán. Reméljük, hogy ez segíteni fogja a magyar NLP közösség céljait.

Eddig mindössze egyetlen hasonló terjedelmű tanulmány készült a magyar nyelvre elérhető nyelvtechnológiáról: 2012-ben az európai META-NET hálózat és partnerintézményei mérték fel az európai nyelvek LT-támogatottságát, és eredményeiket a White Paper sorozatban publikálták. Jelen kötet e módon A magyar nyelv a digitális korban (Simon et al, 2012) utódjának is tekinthető. A Simon és munkatársai által írt áttekintés publikálása óta eltelt majdnem egy évtizedben a nyelvtechnológia forradalmi újításokon esett át: a statisztikai módszereket felváltották a neurálisháló-alapú rendszerek. Míg korábban az NLP-ben szabályalapú módszerekre támaszkodtak a kutatók és fejlesztők, mostanra a legújabb megoldások a mesterséges intelligencia és azon belül a gépi tanulás felől közelítik meg a nyelvfeldolgozási feladatokat. Fontos változás emellett, hogy a nyelvtechnológia vívmányai is egyre szervesebben, egyre fejlettebb formában vannak jelen a mindennapi életünkben. Gondoljunk csak arra, hogy "egyszerűbb", zárt rendszert képező alkalmazási területektől (például a Keleti pályaudvar előzetesen felvett hangfelvételeken (korpuszon) alapuló bemondórendszere) mostanra eljutottunk a nyílt alkalmazásig, például a diktálásig, vagy hogy okostelefonunk vagy tévénk egyes nyelveken már hangvezérelve is működik. Az új technológiák azonban részben új nyelvi adatokat is kívánnak, méghozzá többet, mint korábban bármikor. Ezt a léptékváltást és forradalmi fejlődést, a nyelvmodellek és mesterséges intelligencia alkalmazásán alapuló új korszak kihívásait jelen írásban is érzékeltetni kívánjuk. A továbbiakban a nagyobb erőforrás-kategóriákhoz kapcsolódó, magyar nyelvre kidolgozott adatbázisokat és eszközöket mutatjuk be. Egynyelvű korpuszok esetében több mint 40 adatbázist találtunk. Ezek közül a legnagyobb a több mint 9 milliárd szót tartalmazó, Common Crawl alapú Webcorpus 2.0, amelyet főként nyelvi modellek építésénél használnak. Nagyságban a következő a Magyar Nemzeti Szövegtár 2, amely több mint egy 1 milliárd szónyi gondozott szöveget tartalmaz 6 alkorpuszból, és amiben a határon túli magyar nyelvváltozatok is megjelennek. Az NLP feladatokban kiemelten fontos szerepe van a korpuszok annotációjának és a doménnek is. Magyar nyelvű korpuszoknál főként morfológiai, szintaktikai, illetve néhány esetben egyszerűbb szemantikai annotációk jellemzőek. Speciális annotációra jó példa a NerKor, ami egy 1 millió tokent tartalmazó gold standard névannotált korpusz több doménnel. A doménspecifikus korpuszok tekintetében kiemelkedő a 31 millió tokenes MARCELL korpusz a jogi szakterületről, illetve a BioScope korpusz, amely orvosi szövegeket tartalmaz. Két- és többnyelvű korpuszok esetében több mint 250 adatkészletet találtunk. Ez a nagy szám annak köszönhető, hogy számos nagyszabású nemzetközi projekt tartalmaz magyar nyelvű alkorpuszt is. Ilyenek például az OPUS, az OSCAR vagy a CCMatrix. Bár a legtöbb adatot természetesen angol-magyar nyelvpárra találjuk, az utóbbi években egyre több kisebb nyelvet tartalmazó nyelvpárra is létrejöttek korpuszok, amelyek elengedhetetlenek a fordítóprogramok fejlesztésében.

Az, hogy a nyelvmodellek egyre fontosabb szerepet kapnak a nyelvtechnológiában, erősen meghatározza azt, hogy milyen erőforrásokra van szükség ahhoz, hogy a magyar nyelvnek is megfelelő legyen a technológiai támogatottsága a mesterséges intelligencia korában. Ezt szem előtt tartva a fent bemutatott kategóriákban (ahogy a nyelvtechnológia számos más területén is) egységesen igaz, hogy az eddigieknél több, nagyobb méretű, részletesebben annotált általános és főként doménspecifikus korpuszra van szükség a fejlődéshez.

Az elmúlt néhány évben magyar nyelvre is születtek nyelvmodellek: a HuBERT és a HILBERT, valamint a Hilanco projekt keretében több kísérleti modell is létrejött. Szintén jól állunk a magyar nyelvre kidolgozott elemzőkből illetve elemzőláncokból: a magyarlánc, az e-magyar, a UDpipe és a HuSpaCy megbízható megoldásokat nyújtanak. Ezen a területen is megfigyelhető, és különösen fontos az a tendencia, hogy ipari felhasználást is lehetővé

tevő licencek alatt adják ki az eszközöket. Beszédfeldolgozásban a már említett Keleti pályaudvar bemondója mellett egyre nagyobb teret kapnak a mély tanuláson alapuló technikák, például a Profivox magyar nyelvű szövegfelolvasó alkalmazásában, vagy a Clemvoice illetve a SpeechTex alkalmazásokban. Gépi fordításban szintén a neurális módszer hozza a jobb eredményeket, azonban itt kiemelkedően tudnak teljesíteni az olyan megoldások, amelyek mindezeket nagy pontosságú doménspecifikus adatokkal ötvözik, mint például a Globalese szolgáltatása.

Ahogy láttuk, a több mint 500 elemű magyar NLP erőforrás adatbázis azt mutatja, hogy a nyelvtechnológia számos területén vannak kiemelkedő minőségű korpuszok és eszközök a magyar nyelv technológiai támogatására. Különösen igaz ez például a nyelvi elemzők esetében. Azonban, ahogy a mesterséges intelligencia egyre inkább átszövi a nyelvtechnológiai megoldásokat, úgy változik az is, hogy milyen erőforrások szükségesek ahhoz, hogy a magyar nyelv lépést tudjon tartani a digitalizációval. A nyelvi modellek tanításához egyrészt hatalmas mennyiségű adatra van szükség, másrészt pedig a magasabb nyelvi szinteken annotált, illetve doménspecifikus korpuszok is elengedhetetlenek a fejlesztéshez – és ez a nyelvtechnológia legtöbb területére igaz. Számos kiemelkedő megoldás jelent meg az utóbbi néhány évben a szövegfeldolgozástól a gépi fordításig, azonban még vannak olyan területek, például az összetettebb chatbotok esetében, amelyek nem lefedettek a magyarra. Ugyanakkor a magyar erőforrások esetében is körvonalazódik az a trend, hogy egyre szélesebb kör számára, akár ipari alkalmazásra is elérhetővé teszik az adatbázisokat illetve eszközöket.

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 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.[2] 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 Hungarian Language in the Digital Age

Hungarian, spoken by 13-14 million people worldwide, is the official language of Hungary and a few Hungarian-majority regions and municipalities in Serbia and Slovenia. 9.8 million speakers live in Hungary and a further 2.5 million speakers use Hungarian as a recognised minority language in neighbouring countries that once belonged to Hungary. An additional

[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] https://european-language-equality.eu

one million speakers, most of whom emigrated from Hungary, can be found in Western and West-Central Europe, Southern and Northern America, Australia, and Israel (Fenyvesi, 2005). Albeit the geographical dispersion of speakers, all dialects of Hungarian are mutually intelligible (Kenesei et al., 2002).

Hungarian is the largest Uralic language spoken today. It belongs to the Finno-Ugric subgroup of the Uralic language family, in which its most closely related languages are two critically endangered Ugric languages of western Siberia, Khanty and Mansi. Looking further, we find more distantly related Finnic languages in Northeastern Europe, most importantly, Finnish and Estonian, with a total number of speakers below 7 million combined. Hungary and Hungarians are thus a language island amongst surrounding Indo-European languages of Central Europe, and they are far removed even from the most immediately related languages.

This has important implications for Hungarian language technology: unlike very closely related languages found amongst Indo-European languages with an abundance of similarities between them, Hungarian LT cannot draw much support from the technological development of its closest relatives. Even though Hungarian lacks grammatical gender, developers of Hungarian LT face problems such as the extensive case system and agglutination in the language as nominals inflect for number, case, and person, and verbs inflect for person, number, tense, and mood. The Hungarian case system is particularly complex compared to Indo-European languages. As noted by Thomason, while no modern Indo-European language has more than seven cases, for Hungarian, the analyses range between 17 and 27 (Thomason, 2005). Moreover, there are notable differences between Hungarian and Indo-European languages in their sound systems. Hungarian has short and long vowels (including front rounded vowel phonemes), vowel harmony, fixed word-initial stress and more palatal consonants than Indo-European languages. Some of these differences are reflected in the Hungarian writing system as well, which uses an extended version of the Latin script. Long vowels are marked with an accent (á, é, í, ó, ő, ú, ű), and palatal consonants are written with 'y': ny stands for /ɲ/, gy for /ɟ/, and ty for /c/. There is one three-digit consonant, dzs, but it is only used in words adopted from foreign languages. The current writing system is used since the publication of the *Magyar helyesirás' és szóragasztás' főbb szabályai* in 1832. Overall, including the rarely used q, w, x, and y, the Hungarian alphabet has 44 letters.

Hungarian has a notable presence online. In the Hungarian population aged between 16 and 74, 88% of the households had access to the internet and 79% of the population admitted to using the internet on a daily basis in 2020 (Hivatal, 2020). In November 2021, there were over 845 thousand registered .hu domains, which is the country-code top-level domain (ccTD) of Hungary (Testület, 2021). However, it is important to note that much of the Hungarian-language content on the internet does not actually belong to .hu domains. For example, social media sites (used by 74% of the Hungarian population (Hivatal, 2020) such as Facebook or YouTube, or forums and online newspapers of Hungarian minority communities use the ccTD of their respective countries, .com, .sk, .rs, .ro, etc. Furthermore, the line between Hungarian and non-Hungarian online content cannot be clearly drawn either – a large number of Hungarian internet users participate in global trends and jokes on social media by re-using English phrases and expressions. The resulting mix of Hungarian and English is commonly called Hunglish.

Most of the Hungarian-specific LT resources are developed either in Hungary or as part of large, multilingual Pan-European initiatives. The language variant these resources represent is almost exclusively standard Hungarian. Even in the case of corpora, most of the material that creators include come from within Hungary. One large-scale counterexample that specifically aimed to include regional varieties, i. e. texts from Hungarian communities in neighbouring countries, is the Hungarian National Corpus 2 of 1.5 billion words.

The number of Hungarian speakers has been steadily declining for over 30 years. This is due to several factors. Most importantly, the population of Hungary (where 99% of the total

population speaks Hungarian as their native language) has been declining since 1981 (Hivatal, 2020), and Hungarian minority communities around the world have gradually assimilated into the majority societies of their countries. According to Csete and colleagues' calculations, between 1991 and 2011 alone, the number of Hungarians in neighbouring countries has decreased by almost 600 thousand, from 2.76 million to 2.19 million (Csete et al., 2010). The continuation of this trend indicated a 30% decline from 1991 to 2021 (Csete et al., 2010). In an effort to compensate for these effects and to secure the survival of Hungarian minority communities, Hungarian-language schools operate in Romania, Serbia, Ukraine, Slovakia; and Hungarian bilingual education is available in Slovenia (Csete et al., 2010).

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 because understanding language is a very complex task; it requires understanding the relationship between words, used in different types of texts (genres) and in different situational contexts, as well as to what the words refer to. To understand these relationships, one needs to have textual, contextual and what is often called "world knowledge". Depending on text and context, messages containing similar information can be lexicalised in different ways and create different socially purposeful meanings.

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 Artificial Intelligence (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 (ML), 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 revolutionizing the way in which LT tasks are

[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).

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 Artificial Intelligence (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, 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.
- **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.
- **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.[4]

[4] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is

4 Language Technology for Hungarian

The number of application areas of Hungarian NLP technologies has largely increased since the publication of the META-NET White Paper, and it has definitely outgrown the realm of academic research. Thus, compiling a new, detailed account of LT resources and tools is a necessary step to further aid technological adaptation, and to determine directions for research. Moreover, as neural language models (LMs) have become the leading approach in every subfield of NLP, we put a special focus on LMs throughout our analysis. Most of the datasets and tools in the list can be used free of charge for research and education, and some are also available for commercial purposes under different conditions. Although in the past only a few resources were explicitly licensed, thus causing confusion, nowadays there is a shift towards using standard licensing (e. g. CC-BY-SA-4.0). There is now also a growing trend for using open data wherever possible.

4.1 Language Data

Monolingual corpora

There are more than 40 monolingual text corpora for Hungarian. Amongst these, the largest one is the non-domain specific Hungarian Webcorpus 2.0 (Nemeskey, 2020a) with over 9 billion words, built from Common Crawl and produced primarily for the training of language models. The second largest (and most commonly used) corpus compiled for Hungarian is the Hungarian National Corpus 2 (HNC2; Oravecz et al. (2014)) of more than 1 billion words. Texts of the HNC2 belong to six subcorpora: newspaper, literature, science, official, personal, and transcripts of spoken language. Importantly, the corpus contains linguistic data from Hungarian-speaking minorities of neighbouring countries besides standard Hungarian. HNC2 can be queried through the corpus' online interface.

Currently, most corpora available for Hungarian are only annotated for 'lower' levels of language, such as syntactic and some basic semantic properties. The 82,000-sentence (1.2 million words) Szeged Treebank is the largest fully manually annotated treebank of the Hungarian language. As for specific annotations, NerKor (Simon and Vadász, 2021) is a gold standard named entity annotated corpus containing 1 million tokens; and KorKorpusz contains annotations for coreference and anaphora (Vadász, 2020) in 1,400 sentences. Examples of corpora with higher level linguistic annotation are OpinHuBank, a 10,000-sentence human-annotated corpus compiled to aid the research of opinion mining and sentiment analysis (Miháltz, 2013), or HuSent, a deeply annotated Hungarian sentiment corpus that contains 17,000 sentences of customer reviews of different products (Szabó et al., 2016).

The newly released HuLu (Hungarian Language Understanding Evaluation Benchmark Kit) corpus (Ligeti-Nagy et al., 2022) is being developed as the Hungarian version of the GLUE and SuperGLUE benchmark databases which are English standards for benchmarking. Just as GLUE and SuperGLUE, HuLu can be used primarily for the evaluation and analysis of natural language understanding (NLU) systems.

Overall, there are very few domain-specific monolingual corpora currently available for Hungarian, and those that exist are mostly from the legal domain (e. g. MARCELL (Váradi et al., 2020), Miskolc Legal Corpus (Vincze, 2018), the Hungarian subcorpus of EuroParl (Koehn, 2005), etc.). There are even fewer datasets from the health domain (e. g. BioScope (Szarvas et al., 2008)) despite its importance. Other domains such as customer service or social media texts are either too small (they are minor subcorpora in larger datasets) or are almost

anticipated to grow at an annual rate of 18,4% from 2020 to 2028 (https://tinyurl.com/2p9ed6tp). 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).

entirely missing. This is an obstacle in the development of chatbots, automated customer service systems and applications for filtering fake news.

Bi- and multilingual text corpora

Multilingual textual data containing Hungarian are abundant with almost 250 datasets. For example, an important resource is the OSCAR corpus (Ortiz Suárez et al., 2019) compiled from the CommonCrawl corpus for 166 languages, or Plaintext Wikipedia dumps 2018 comprising 297 Wikipedias.[5] OPUS (Tiedemann (2012)) is a growing collection of crawled, translated and sentence-aligned open source corpora. Here the English-Hungarian language pair contains about 65.5 million segments and 854 million Hungarian tokens. A large collection of Hungarian comparable corpora is also available, or currently under construction, for example, in the framework of the ParlaMint project of CLARIN.

There are also several domain-specific multilingual parallel corpora hosted at the the EU Science Hub for all the EU languages including Hungarian (see for example JRC-Acquis[6] or the EAC Translation Memory,[7] for a detailed description see Steinberger et al. (2014)).

While datasets like OSCAR and OPUS facilitate the research on cross-lingual transfer learning, there is still a huge need for parallel corpora for neural machine translation (NMT). At large, general-domain parallel corpora for well-resourced languages other than English are extremely scarce, thus creating a bottleneck for NMT. Thanks to the CCMatrix (Schwenk et al. (2021)), a huge amount of corpora has become available recently for additional language pairs. This allows the creation of direct translation models and eliminates the need to use English as an intermediate language.

In addition to the large-scale multilingual, automatically compiled corpora, several datasets have been created in Hungarian LT centres as well. Among these, the most prominent is Hunglish (Varga et al., 2005), a freely available sentence-aligned Hungarian-English parallel corpus of about 120 million words in 4 million sentence-pairs. One outstanding exception for the pair of Hungarian and another well-resourced language is HunOr (Szabó et al., 2012), a multi-domain Hungarian–Russian parallel corpus containing approximately 800,000 words.

Multimodal corpora (audio, video)

The number of multimodal corpora for Hungarian is quite low, with the most common form being an audio dataset backed with transcripts (e. g. BEA (the Hungarian Spontaneous Speech Database (Gósy et al., 2012)); the Budapest Sociolinguistic Interview (Kontra and Váradi, 1997)). Multinational projects have also aimed at collecting spoken data from phone speech or reading, see for example MaSS – Multilingual corpus of Sentence-aligned Spoken utterances from the Bible (Boito et al., 2019) or CSS10, a collection of single-speaker speech datasets of 10 languages including Hungarian (Park and Mulc, 2019)). Importantly, as there are no publicly available domain-specific multimodal datasets of considerable sizes in Hungarian, R&D projects tend to compile their own resources (Mihajlik et al., 2021) to train and evaluate speech processing systems.

As for multimodal annotated video corpora, the 50-hour fully transcribed and richly annotated HuComTech corpus (Hunyadi et al., 2018) (annotated for both visual and auditory properties such as facial expressions, eyebrow movement, gaze, headshift, shape of hands, gestures, posture, emotions, discourse, prosody) is a unique achievement not just for Hungarian but by international standards.

[5] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2735
[6] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[7] https://ec.europa.eu/jrc/en/language-technologies/eac-translation-memory

Lexical/conceptual resources

Lexical/conceptual resources in Hungarian mostly include datasets of the structure and behaviour of Hungarian verbs. Some examples of these are the following: the Mazsola database of verb frames (Sass, 2015), the Tádé frequency list of verbal argument structures (Kornai et al., 2016), the PrevCons list of hapaxes of verbs with verbal prefixes and their verb frames (Kalivoda, 2021), and the PrevLex list of Hungarian phrasal verbs; (Kalivoda, 2018). An important ontology is the Hungarian WordNet (Miháltz et al., 2008) that includes business and legal domains and contains over 42,000 synsets.

Neural Language Models

Following the international R&D trends of the past few years, there has been a huge growth in developing LM solutions for Hungarian. As BERT became a standard in NLP (e. g. (Rogers et al., 2020)), a number of LMs with BERT architecture were trained for Hungarian, first and foremost HuBERT (Nemeskey, 2020a), HILBERT (Feldmann et al., 2021), emBERT (Nemeskey, 2020b), and a couple of experimental models developed by the HILANCO consortium.[8] Besides BERT, models with other architectures are being continuously adapted to Hungarian. Moreover, there are multilingual LMs including Hungarian, for instance, mBERT (Devlin et al., 2018) pretrained on the Wikipedias of 104 languages.

4.2 Language Technologies and Tools

Text Analysis

Solutions for the most common tasks in text analysis are currently available in state-of-the-art NLP tools and pipelines that perform highly accurate linguistic analysis in Hungarian. While UDPipe (Straka and Straková, 2017), Stanza (Qi et al., 2020) and HuSpaCy (Orosz et al., 2022) are neural network models added to a multilingual framework, there are also two toolchains built specifically for Hungarian, e-magyar (Váradi et al., 2018) (Simon et al., 2020)) and Magyarlánc (Zsibrita et al., 2013). The recently upgraded HuSpaCy provides reliable industrial-grade Hungarian language processing facilities, and covers tokenisation, sentence splitting, PoS tagging, lemmatisation, dependency parsing, named entity recognition and word embedding representation. Despite the recent progress of LMs, there is a huge need for preprocessing pipelines especially in commercial application.

Additionally, the Trendminer Hungarian Processing Pipeline (Miháltz et al., 2015) performs linguistic analysis of social media texts by adopting existing toolchains; and NooJ, a finite-state transducer with a Hungarian module (Váradi and Gábor, 2004), is still used to carry out higher level analysis of texts (e. g. psychological investigations of political attitudes) by some research groups (e. g. Ilg (2021)). Unfortunately, these two tools have not been updated recently. There are also some text analysis toolkits available in Hungarian developed by industrial stakeholders, e. g., Neticle's media monitoring system.[9]

Speech Processing

Although there are numerous multilingual speech processing tools covering Hungarian, only a few Hungarian-specific applications are available.[10] Just like in many other subfields of NLP, the most popular text-to-speech (TTS) paradigm is to substitute the whole chain by deep neural networks (DNNs) (Ning et al., 2019). Parallel to this line of development, constantly

[8] https://hilanco.github.io
[9] https://neticle.com/mediaintelligence/hu
[10] We thank Péter Mihajlik for his valuable comments on this section.

revised versions of the Profivox system (Olaszy et al., 2000), developed by TMIT BME, have been providing TTS solutions for Hungarian for over twenty years now, ranging from dyad-based systems to state of the art DNN models. Similarly, in ASR research and development the DNN approach has become prominent (Mihajlik et al., 2021). As for commercial applications see, for instance, Clementine's Clemvoice[11] that provides services including speech processing, or SpeechTex[12] specialising in TTS for the legal domain.

In speech processing, just like at every other part of LT where LMs are used, there is a lack of computational and speech resources, i. e. competitive GPU-grids and high variability natural speech recordings, that hinders development in the fields of TTS and ASR.

Translation Technologies

While a number of approaches and architectures have been proposed and tested over the years (e. g., the pattern-based MT system MetaMorpho (Prószéky and Tihanyi, 2002) or statistical systems (SMT) (Laki et al., 2013)) in Hungarian, neural machine translation (NMT) has become the leading paradigm for MT in the last couple of years. The state-of-the-art NMT system is implemented by Laki and Yang (2022). To carry out high-performance NMT, however, the collection of high quality parallel language data both from general and specific domains is essential. The Hungarian NMT provider Globalese[13] does this by enabling human translators to train the company's NMT engines based on their own parallel data.

Language Generation and Summarisation

Although there are some corporate solutions covering Hungarian (e. g. IntelliDockers engines or SAS), we are not aware of any summarisation tool specifically developed for Hungarian. As first steps towards such a tool, Yang et al. built the first extractive (Yang et al., 2020a) and the first abstractive (Yang et al., 2021) summarisation tools based on Hungarian-specific transformer models. Unfortunately, there are no publicly available summarisation datasets for Hungarian either – in the research of Yang et al. (2021), online news articles and their lead texts were used (published by HVG and index.hu) but this dataset is not publicly available due to legal reasons. Yang (2022) has built the first GPT-2 model (with a news and a poem generator) for Hungarian.

Human-Computer Interaction

There is a growing demand for technological solutions to human-computer interaction. Chatbots and simple, task-based systems are increasingly used, for example, the commercial chatbots developed by RoboRobo are claimed to have had more than one million users so far.[14] Another example is the Hun-appointment-chatbot[15] for appointment bookings. However, systems that can carry out more open-ended conversations in Hungarian are not yet available.

Information retrieval

In the last number of years there have been several initiatives for creating solutions for information retrieval. In the field of web crawling Hungarian Webcorpus 2.0 (Nemeskey, 2020a)

[11] https://clementine.hu/megoldasok/ugyfelszolgalat/clemvoice
[12] https://speechtex.com
[13] https://www.globalese-mt.com
[14] https://roborobo.hu/hu
[15] https://github.com/szegedai/hun-appointment-chatbot

is the largest web corpus for Hungarian. Indig et al. (2020) built a middle-sized corpus using targeted web crawling. The first vector space model (Novák et al., 2017) was also built with a searchable online interface for Hungarian. Yang et al. (2020b) built a text classification, tag recommendation tool for news articles. Osváth et al. (2021) are building annotated corpora and neural models with topic modelling and sentiment analysis to extract patient health care experiences from online fora. Laki and Yang (2021) have built various neural sentiment analysis models for Hungarian.

4.3 Projects, Initiatives, Stakeholders

With the growth of the role of AI in several fields, numerous national programs and umbrella organisations have been founded recently. The two most prominent organisations in Hungary are the Artificial Intelligence National Laboratory and the Artificial Intelligence Coalition. Their goals include facilitating cooperation and communication between research centres, universities and industrial AI developers; and, eventually, to strengthen the position of Hungarian AI internationally.

In the last five years, one of the most prominent projects carried out in the cooperation of several leading Hungarian R&D centres was e-magyar, a state-of-the-art modular toolchain for the Hungarian language, now available for researchers, developers, and for the general public. As for international projects, there have been three large initiatives of regional cooperation recently, namely, ELEXIS, MARCELL, and CURLICAT, all under the coordination of the Hungarian Research Centre for Linguistics.

Regarding industry, there is a growing number of companies offering top quality LT tools and/or services for the Hungarian language ranging from chatbots to automated translation and to more general text analytics solutions.

5 Cross-Language Comparison

The LT field[16] 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[17] 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)

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

[17] 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.

  - 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.)

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[18]

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

[18] 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.

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[19] 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.[20]

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 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,[21] 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

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

[20] 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.

[21] 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.

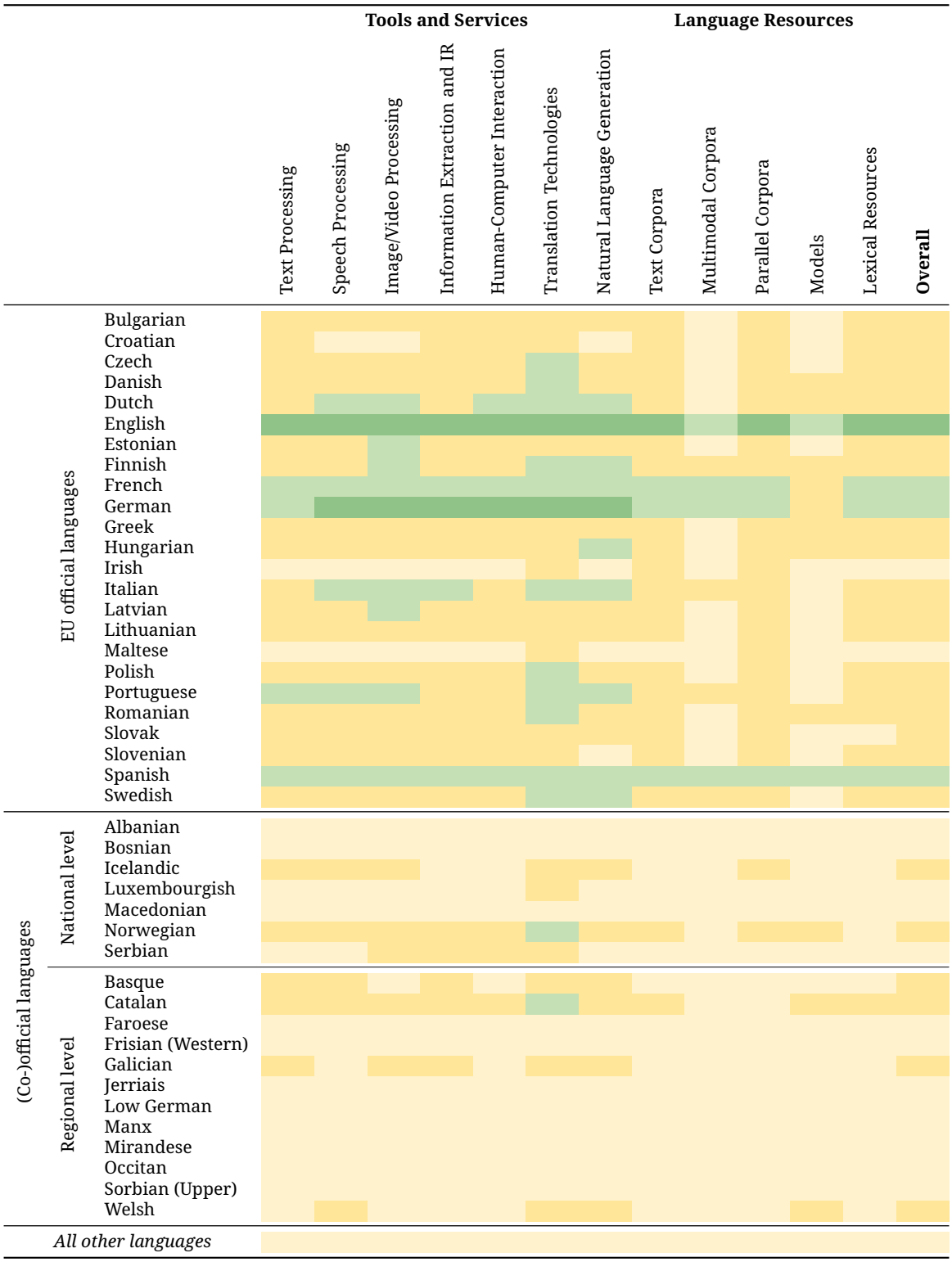

Table 1. State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

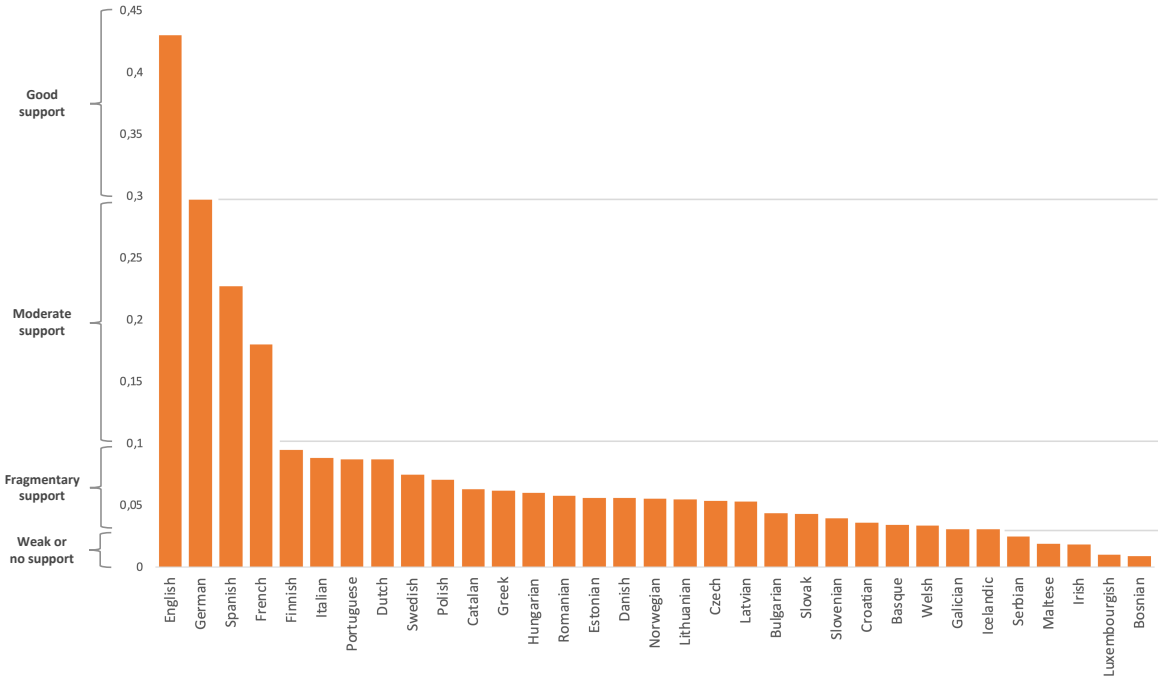

Figure 1. Overall state of technology support for selected European languages (2022)

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 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

This report gives a qualitative overview of the database on the resources currently available in Hungarian NLP. The emergence of neural technologies has massively reshaped how language data is used in a uniform way in most subfields of NLP. As we have seen in examples ranging from speech processing to summarisation and to machine translation, in several areas state-of-the-art results can only be achieved through drawing in 'an unimaginably large amount of data' compared to previous standards. This poses an additional challenge for researchers, as many times they not only need to develop technological solutions but also to find and create their own textual resources. Moreover, although there is clearly room to expand, Hungarian as a medium-sized language is, by default, in a disadvantaged position due to its size. This means that the smaller number of total speakers and platforms generate less harvestable data than what is available for languages like English or German.

Secondly, although plenty of monolingual corpora were compiled in the past years, there is an ever-growing need for novel datasets for fine-tuning, testing and benchmarking. These datasets should contain high quality annotation especially for the higher linguistic levels closer to NLU, covering a wide variety of specific domains. Due to their importance the automatic generation of such resources should be considered as well. There is also a need for more domain-specific and general domain parallel corpora as these are still a key prerequisite for machine translation.

Thanks to the efforts of the last decade, there are now multiple toolchains performing good-quality linguistic analysis. At the same time, more intricate tasks are still to be covered, as tools aiming at higher level analysis of texts are outdated. Processing solutions for social media texts should also be expanded. Human-computer interaction is a sub-field that appears to be of utmost importance at present, however, complex conversational agents are not present for Hungarian as of now, so improvement in this area is also essential.

In terms of supporting R&D, the most positive initiative of recent years is that several umbrella organisations have been established to support NLP in Hungary, to foster cooperation between the most important research centres, and to facilitate the dialogue between R&D and industry. Another positive change underway is the use of standard licenses in order to make resources open-source.

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.

Marcely Zanon Boito, William N. Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *CoRR*, abs/1907.12895, 2019. URL http://arxiv.org/abs/1907.12895.

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

Örs Csete, Attila Papp Z., and János Seténvi. Kárpát medencei magyar oktatás az ezredfordulón. In Ablonczy Balázs, Bárdi Nándor, and Bitskey Botond, editors, *Határon túli magyarság a 21. században*, pages 125–165. Köztársasági Elnöki Hivatal, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Ádám Feldmann, Róbert Hajdu, Balázs Indig, Bálint Sass, Márton Makrai, Iván Mittelhocz, Dávid Halász, Győző Yang Zijian, and Tamás Váradi. Hilbert, magyar nyelvű bert-large modell tanítása felhő környezetben. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 29–36. Szegedi Tudományegyetem TTIK, Informatikai Intézet, 2021.

Anna Fenyvesi. *Hungarian language contact outside Hungary: Studies on Hungarian as a minority language*. John Benjamins Publishing, 2005.

Mária Gósy, Dorottya Gyarmathy, Viktória Horváth, Tekla Etelka Gráczi, András Beke, Tilda Neuberger, and Péter Nikléczy. BEA: Beszélt nyelvi adatbázis . In Gósy Mária, editor, *Beszéd, adatbázis, kutatások*, pages 9–24. Akadémiai Kiadó, 2012.

Központi Statisztikai Hivatal. A háztartások információs- és kommunikációseszköz-használatának főbb jellemzői. https://www.ksh.hu/docs/hun/xftp/idoszaki/ikt/2020/01/index.html, 2020. Accessed: 2021-09-01.

László Hunyadi, Tamás Váradi, György Kovács, István Szekrényes, Hermina Kiss, and Karolina Takács. Human-human, human-machine communication: on the hucomtech multimodal corpus. In Inguna Skadina and Maria Eskevich, editors, *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, pages 56–65. Linköping University Electronic Press, Linköpings universitet, 2018.

Barbara Ilg. The representation of trianon trauma as a chosen trauma in political newspapers (1920-2010) in hungary. *Corvinus Journal of Sociology and Social Policy*, 12:51–93, 2021.

Balázs Indig, Árpád Knap, Zsófia Sárközi-Lindner, Mária Timári, and Gábor Palkó. The ELTE.DH pilot corpus – creating a handcrafted Gigaword web corpus with metadata. In *Proceedings of the 12th Web as Corpus Workshop*, pages 33–41, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-68-9. URL https://aclanthology.org/2020.wac-1.5.

Ágnes Kalivoda. Véges erőforrás végtelen sok igekötős igére. In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, pages 331–344. Szegedi Tudományegyetem TTIK Informatikai Intézet, 2018.

Ágnes Kalivoda. Az igekötők produktív kapcsolódási mintái. *Argumentum*, 17:56–82, 10 2021.

István Kenesei, Robert M. Vágó, and Anna Fenyvesi. *Hungarian*. Routledge, 2002.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

Miklós Kontra and Tamás Váradi. *The Budapest Sociolinguistic Interview: Version 3.* Linguistics Institute, Hungarian Academy of Sciences, 1997.

András Kornai, Dávid Márk Nemeskey, and Gábor Recski. Detecting optional arguments of verbs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2815–2818, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1448.

László Laki, Attila Novák, and Borbála Siklósi. English to Hungarian morpheme-based statistical machine translation system with reordering rules. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 42–50, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2808.

László János Laki and Zijian Győző Yang. Improving performance of sentence-level sentiment analysis with data augmentation methods. In *Proceedings of the 12th IEEE International Conference on Cognitive Infocommunications*, pages 417–422, 2021.

László János Laki and Zijian Győző Yang. Jobban fordítunk magyarra, mint a Google! In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2022)*, pages 357–374, Szeged, 2022.

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, Kinga Jelencsik-Mátyus, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. Hulu: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2022)*, pages 431–448, Szeged, 2022.

Péter Mihajlik, András Balogh, Balázs Tarján, and Tibor Fegyó. End-to-end és hibrid mélyneuronháló alapú gépi leiratozás magyar nyelvű telefonos ügyfélszolgálati beszélgetésekre. In *Magyar Számítógépes Nyelvészeti Konferencia*, pages 139–145. Szegedi Tudományegyetem TTIK Informatikai Intézet, 2021.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János A. Csirik, Gábor Prószéky, and Tamás Váradi. Methods and results of the hungarian wordnet project. In *Proceedings of the Fourth Global WordNet Conference GWC*, pages 310–320, 2008.

Márton Miháltz, Tamás Váradi, István Csertő, Éva Fülöp, Tibor Pólya, and Pál Kővágó. Beyond sentiment: Social psychological analysis of political Facebook comments in Hungary. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–133, Lisboa, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2918. URL https://aclanthology.org/W15-2918.

Márton Miháltz. Opinhubank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In *Magyar Számítógépes Nyelvészeti Konferencia*, pages 343–345. Szegedi Tudományegyetem TTIK Informatikai Intézet, 2013.

Dávid Márk Nemeskey. *Natural Language Processing Methods for Language Modeling*. PhD thesis, Eötvös Loránd University, 2020a.

Dávid Márk Nemeskey. Egy `emBERT` próbáló feladat. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020)*, pages 409–418, Szeged, 2020b.

Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19), 2019. ISSN 2076-3417. doi: 10.3390/app9194050. URL https://www.mdpi.com/2076-3417/9/19/4050.

Attila Novák, Borbala Novák, and Nóra Wenszky. Szóbeágyazási modellek vizualizációjára és böngészésére szolgáló webes felület. In Veronika Vincze, editor, *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, Szeged, 2017. Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Gábor Olaszy, Géza Németh, Péter Olaszi, and Géza Kiss. Profivox: a legkorszerűbb hazai beszédszintetizátor. In *Beszédkutatás*, pages 167–179, 2000.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. The hungarian gigaword corpus. In *Proceedings of LREC*, pages 1719–1723, 2014.

György Orosz, Zsolt Szántó, Péter Berkecz, Gergő Szabó, and Richárd Farkas. HuSpaCy: an industrial-strength Hungarian natural language processing toolkit, 2022.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi, editors, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom, July 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/IDS-PUB-9021. URL https://hal.inria.fr/hal-02148693/file/Asynchronous_Pipeline_for_Processing_Huge_Corpora_on_Medium_to_Low_Resource_Infrastructures.pdf.

Mátyás Osváth, Zijian Győző Yang, and Karolina Kósa. Analysing reactions to patient health care experiences via sentiment analysis and bert topic modeling. In *Proceedings of the 12th IEEE International Conference on Cognitive Infocommunications*, pages 411–416, 2021.

Kyubyong Park and Thomas Mulc. CSS10: A collection of single speaker speech datasets for 10 languages. *CoRR*, abs/1903.11269, 2019. URL http://arxiv.org/abs/1903.11269.

Gábor Prószéky and László Tihanyi. A pattern-based machine translation system. In *Proceedings of the 24th Translating and the Computer Conference*, pages 19–24, 2002.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14. URL https://aclanthology.org/2020.acl-demos.14.

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.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327, 2020. URL https://arxiv.org/abs/2002.12327.

Bálint Sass. 28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet. In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2015)*, pages 303–308, Szeged, 2015.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL https://aclanthology.org/2021.acl-long.507.

Eszter Simon and Noémi Vadász. Introducing nytk-nerkor, A gold standard hungarian named entity annotated corpus. In Kamil Ekstein, Frantisek Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 222–234. Springer, 2021. doi: 10.1007/978-3-030-83527-9_19.

Eszter Simon, Piroska Lendvai, Géza Németh, Gábor Olaszy, and Klára Vicsi. *A magyar nyelv a digitális korban – The Hungarian 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/hungarian. Georg Rehm and Hans Uszkoreit (series editors).

Eszter Simon, Balázs Indig, Ágnes Kalivoda, Iván Mittelholcz, Bálint Sass, and Noémi Vadász. Újabb fejlemények az `e-magyar` háza táján. In Gábor Berend, Gábor Gosztolya, and Veronika Vincze, editors, *XVI. Magyar Számítógépes Nyelvészeti Konferencia*, pages 29–42, Szeged, 2020. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. An overview of the european union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707, dec 2014. ISSN 1574-020X. doi: 10.1007/s10579-014-9277-0. URL https://doi.org/10.1007/s10579-014-9277-0.

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. doi: 10.18653/v1/K17-3009. URL https://aclanthology.org/K17-3009.

Martina Katalin Szabó, Veronika Vincze, and István Nagy T. A hungarian-russian parallel corpus. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2453–2458, Istanbul, 2012. European Language Resources Association (ELRA).

Martina Katalin Szabó, Veronika Vincze, Katalin Ilona Simkó, viktor Varga, and Viktor Hangya. A hungarian sentiment corpus manually annotated at aspect level. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2873–2878, Portoroz, 2016. European Language Resources Association (ELRA).

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://aclanthology.org/W08-0606.

Magyarországi Internetszolgáltatók Tanácsa Tudományos Testület. Magyarországi internet szolgáltatók tanácsa tudományos egyesület. https://info.domain.hu/stats/hu, 2021. Accessed: 2021-09-01.

Sarah G. Thomason. Typological and theoretical aspects of hungarian in contact with other languages. In *Hungarian language contact outside Hungary*, pages 11–28, Amsterdam – Philadelphia, 2005. John Benjamins.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

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.

Noémi Vadász. KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése. In Gábor Berend, Gábor Gosztolya, and Veronika Vincze, editors, *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020)*, pages 141–154, Szeged, 2020. Szegedi Tudományegyetem, TTIK, Informatikai Intézet.

Tamás Váradi, Eszter Simon, Bálint Sass, Iván Mittelholcz, Attila Novák, Balázs Indig, Richárd Farkas, and Veronika Vincze. E-magyar – a digital language processing system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1208.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. The MARCELL legislative corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.464.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596, Szeged, 2005. Szegedi Tudományegyetem, TTIK, Informatikai Intézet.

Veronika Vincze. A miskolc jogi korpusz nyelvi jellemzői. In Miklós Szabó, editor, *A törvény szavai*, pages 9–36, Miskolc, 2018. Miskolci Egyetem - MAB.

Tamás Váradi and Kata Gábor. A magyar intex fejlesztéséről. In Zoltán Alexin and Dóra Csendes, editors, *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004)*, pages 3–10, Szeged, 2004. Szegedi Tudományegyetem.

Zijian Győző Yang, Attila Perlaki, and László János Laki. Automatikus összefoglaló-generálás magyar nyelvre bert modellel. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia*, pages 343–354, Szeged, Magyarország, 2020a. Szegedi Tudományegyetem, Informatikai Intézet.

Zijian Győző Yang, Ádám Agócs, Gábor Kusper, and Tamás Váradi. Abstractive text summarization for hungarian. *Annales Mathematicae et Informaticae*, 53:299–316, 2021.

Zijian Győző Yang. ”az invazív medvék nem tolerálják a suzukis agressziót” – Magyar GPT2 kísérleti modell. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2022)*, pages 463–476, Szeged, 2022.

Zijian Győző Yang, Attila Novák, and László János Laki. Automatic tag recommendation for news articles. In *Proceedings of the 11th International Conference on Applied Informatics*, pages 442–451, Eger, 2020b. CEUR Workshop Proceedings.

János Zsibrita, Veronika Vincze, and Richárd Farkas. magyarlanc: A tool for morphological and dependency parsing of Hungarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 763–771, Hissar, Bulgaria, September 2013. INCOMA Ltd. Shoumen, BULGARIA. URL https://aclanthology.org/R13-1099.