EUROPEAN
LANGUAGE
EQUALITY

D1.7

Report on the Croatian Language

Author	Marko Tadić
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	D1.7
Deliverable title	Report on the Croatian Language
Type	Report
Number of pages	30
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	Marko Tadić
Reviewers	Tea Vojtěchová, Annika Grützner-Zahn
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2022 ELE Consortium

Consortium

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

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
CALL	Computer Assisted Language Learning
CEF AT	Connecting Europe Facility, Automated Translation
CLARIN	Common Language Resources and Technology Infrastructure
DLE	Digital Language Equality
EC	European Commission
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)
ELRC	European Language Resource Coordination
HPC	High-Performance Computing
LPC	Language Processing Chain
LPP	Language Processing Pipeline
LR	Language Resources/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NLG	Natural Language Generation
UD	Universal Dependencies

Abstract

The *Language Report on Croatian* is part of a series of reports on the level of support the European languages receive through technology. In this report, provided within the European Language Equality project, the developments in the last ten years and since the last such report (Tadić et al., 2012) are presented. In Section 2.1 the general information about historical, genealogical, typological and structural features of the Croatian language, its official status and usage, as well as literacy level are described. Section 2.2 focuses on the role and presence of the Croatian language in the digital sphere, starting with general statistics on the households equipped with ICT, up to the presence of Croatian online in different formats: WWW, Wikipedia, social media, Machine Translation, localisations of major operating systems and software packages etc.

After the general description of Language Technology in Section 3, a precise snapshot of the state of language technological support for Croatian is presented in Section 4. This information is segmented in three Subsections 4.1 Language Data, 4.2 Language Technologies and Tools and 4.3 Projects, Initiatives, Stakeholders. In Section 4.1 a list of monolingual, bilingual and multilingual corpora collected in the last period is given. Also, the situation with the lexical resources as well as computational grammars and language models is described. In Section 4.2 the progress in development of different tools and technologies for processing and using Croatian is presented, particularly the existing Language Processing Chains (LPC) or Language Processing Pipelines (LPP) and the inclusion or exclusion of Croatian in different available NLP platforms. In the same section the developments in translation technologies and Computer Assisted Language Learning (CALL) are also presented. The section finishes with the warning for serious underdevelpment in the field of speech processing. In Section 4.3 previous national and EU funded projects are presented as well as main players in the field of Language Technologies in Croatia. The role of full membership in CLARIN ERIC is also stressed as one of factors that rises the visibility of LT. The section finishes with the remark on using so-called Serbo-Croatian language resources where it is explained how the usage of such resources for processing Croatian would yield noisy and error prone results.

This report provides also a cross-language comparison in Section 5. Croatian is positioned in the group of languages with moderate technological support with some subfields of weak or no support at all. The report ends with conclusions and projection of plans for further steps in providing the technological support for the Croatian language.

Sažetak

Ovo *Izvješće o hrvatskome jeziku* dio je niza izvješća o razini jezičnotehnološke podrške za europske jezike. U ovom izvješću, koje je izrađeno u okviru projekta Europska jezična ravnopravnost (European Language Equality), prikazan je razvitak u posljednjih deset godina i od posljednjega takvoga izvješća (Tadić et al., 2012). U odjeljku 2.1. opisane su opće informacije o povijesnim, genealoškim, tipološkim i strukturnim obilježjima hrvatskoga jezika, o njegovu službenom statusu i uporabi, kao i o razini pismenosti. Odjeljak 2.2. usmjeren je na ulogu i prisutnost hrvatskoga jezika u digitalnoj sferi, počevši od općih statističkih podataka o broju kućanstava opremljenih IKT-om, do prisutnosti hrvatskoga jezika *on-line* u različitim formatima: WWW, Wikipedia, društveni mediji, strojno prevođenje, lokalizacije glavnih operacijskih sustava i programskih paketa itd.

Nakon općega opisa jezične tehnologije u odjeljku 3., u odjeljku 4. prikazan je precizan prikaz stanja jezičnotehnološke podrške hrvatskomu jeziku. Te su informacije podijeljene u tri pododjeljka 4.1. Podatci o jeziku, 4.2. Jezične tehnologije i alati te 4.3 Projekti, inicijative, dionici. U odjeljku 4.1. dan je popis jednojezičnih, dvojezičnih i višejezičnih korpusa

prikupljenih u posljednjem razdoblju. Također, opisano je stanje s leksičkim resursima te računalnim gramatikama i jezičnim modelima. U odjeljku 4.2. prikazan je napredak u razvoju različitih alata i tehnologija za obradbu i korištenje hrvatskoga jezika, posebno postojećih lanaca za obradbu jezika (Language Processing Chain, LPC) ili nizova za obradbu jezika (Language Processing Pipeline, LPP) te uključivanje i/li isključivanje hrvatskoga jezika u različitim dostupnim platformama za obradbu prirodnoga jezika. U istom poglavlju prikazan je razvoj prevoditeljskih tehnologija i računalno podržanoga učenja jezika (Computer Assisted Language Learning, CALL). Odjeljak završava upozorenjem o ozbiljnome zaostajanju u području računalne obradbe hrvatskoga govora. U odjeljku 4.3. prikazani su dosadašnji nacionalni projekti i projekti koje financira EU, kao i glavni sudionici istraživanja s područja jezičnih tehnologija u Hrvatskoj. Također je naglašena uloga punopravnoga članstva u CLARIN ERIC-u kao jednoga od čimbenika koji jezične tehnologije čini vidljivijima. Poglavlje završava napomenom o korištenju tzv. srpsko-hrvatskih jezičnih resursa i objašnjava kako bi korištenje takvih resursa za obradbu hrvatskoga jezika nesumnjivo rezultiralo nepreciznim i zapravo pogrješnim rezultatima.

U odjeljku 5. ovoga izvješća prikazana je i usporedba među jezicima gdje je u odjeljku 5.1. prikazana metodologija te usporedbe, u odjeljku 5.2. definirani su stupnjevi tehnološke potpore, u odjeljku 5.3. prikazana je uloga Europske jezične mreže (European Language Grid, ELG), a u odjeljku 5.4. prikazani su rezultati i nalazi, te se o njima raspravlja. Hrvatski je jezik smješten u skupinu jezika s umjerenom tehnološkom podrškom s nekim područjima vrlo slabe ili nikakve potpore. Izvješće završava zaključcima i projekcijom planova za daljnje korake u pružanju tehnološke podrške promatranim jezicima.

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 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 – 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.

[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

2 The Croatian Language in the Digital Age

2.1 General Facts

The Croatian language belongs to the West-South Slavic subgroup of the Balto-Slavic branch of the Indo-European linguistic family. Currently, over 5.5 million people speak Croatian as their native language. The Croatian language consists of the dialects and standard national language of the Croats, which is the official language of more than 4 million people in the Republic of Croatia and is, along with Bosnian and Serbian, one of the three official languages in Bosnia and Herzegovina, where it is spoken by about 700,000 people. However, the Croatian language is also spoken by members of national minorities in Croatia as well as by autochthonous Croatian ethnic and linguistic minorities in Serbia, Montenegro, Slovenia, Hungary, Austria, Slovakia and Italy, who either reside upon territories of former Croatian lands or emigrated due to historically conditioned exoduses throughout the centuries. Due to intensive economically and politically conditioned emigration in late 19th and early 20th century and particularly after the Second World War, Croatian is also spoken within the Croatian linguistic community in a number of other European countries and overseas. The largest Croatian economic diaspora is located in Germany, followed by the USA, Canada and Australia, and they also occasionally use the Croatian language.

The official status of the Croatian language in Croatia is defined by the Constitution of the Republic of Croatia, Article 12: "The Croatian language and the Latin script shall be in the official use in the Republic of Croatia. In particular local units, another language and Cyrillic script or some other script may be introduced into the official use together with the Croatian language and Latin script under conditions specified by law." By the law on the national minorities and their rights, when the local unit (county, town or municipality) has members of a minority over 1/3 of total inhabitants, the conditions for the official use of that minority language are met. Croatian is used in entire territory of the Republic of Croatia as well as neighbouring countries and overseas. Since Croatia joined the European Union in 2013, the Croatian language became the 24th official language of the EU.

According to the 2011 census (Census, 2013), Croatia has 4,284,889 inhabitants of which 90.42% are Croats and Croatian is the native language of 95.60% of all residents of the Republic of Croatia.[3]

The dialectal map of Croatia is composed of three dialectal groups: Čakavian, Kajkavian and Štokavian (see Figure 1). Dialects belonging to all three dialectal groups are spoken throughout the Republic of Croatia. All Croatian dialects belong to the Central South Slavic diasystem of the Slavic linguistic branch, and on the South-Slavic territory it comprises part of the dialectal continuum between the Slovenian type in the North-West and the Macedonian-Bulgarian type in the South-East. The names of those dialectal groups are based upon the use of the interrogative pronouns *ča*, *kaj* and *što* 'what' (lat. *quid*). However, between South-Slavic languages, the Croatian language is the only one that encompasses these three dialects, so this classification is relevant only for Croatian and not for other South-Slavic languages.

The history of the Croatian language is attested to by texts written as early as the end of the 10th or the beginning of the 11th century, the period in which the three Croatian dialects (Čakavian, Štokavian, Kajkavian) began to form. All three Croatian dialects played an important part in the formation of the Croatian literary language (various dialectal stylizations) and the moulding of the Croatian linguistic culture that led to the Croatian standard language with a Štokavian foundation. The first clear trends towards the shaping of the Croatian standard language became apparent in the 17th century, when the majority of the Croatian ethnic community – especially after the grammar and other works of Bartol Kašić (1575–1650) and a flourishing of Renaissance and Baroque literature from Štokavian Dubrovnik – recognised

[3] https://www.dzs.hr/Hrv_Eng/publication/2012/SI-1469.pdf (accessed 2021-12-15).

the linguistic structure of the Štokavian dialect as the best starting point for the construction of a supra-regional Croatian literary language. Although the standardisation of the language of the Croats based upon the Štokavian dialect began very early, national linguistic unity was only achieved during the time of the Illyrian national revival (starting in 1835).

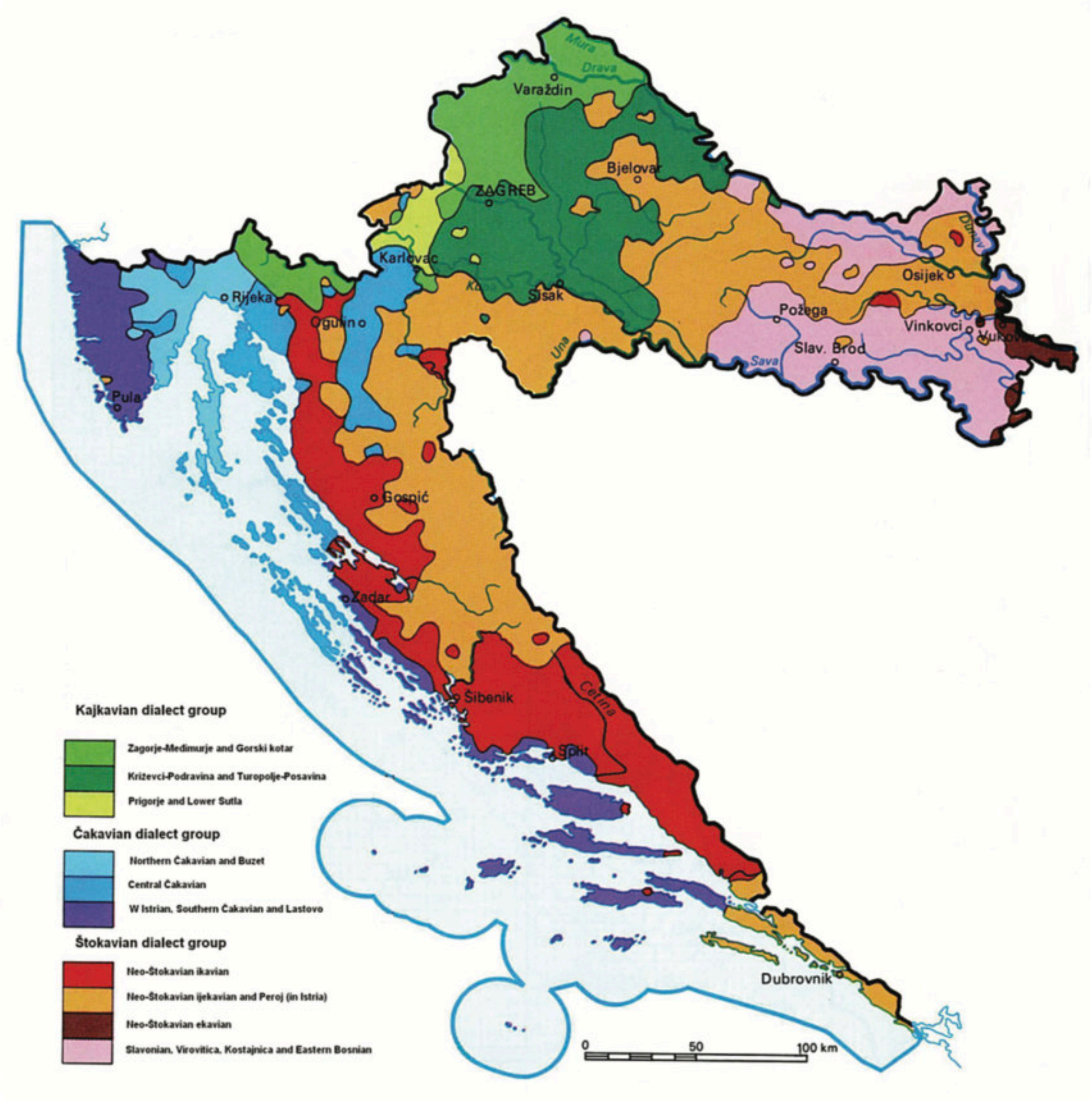

Figure 1: Map of Croatian dialects in the Republic of Croatia (Tadić et al., 2012)

Croatian written culture is marked by the use of three scripts and alphabets (Glagolitic, Cyrillic, Latin), and the Latin script has been the foremost of the three among the Croats since the 16th century. Its usage was neither standardised nor systematised until 1835, when Ljudevit Gaj gave the Croatian Latin alphabet its modern-day form.

Linguistic features of the Croatian language will be described here as stratified to the language levels following (Tadić et al., 2012).

The phoneme inventory of the Croatian standard language consists of 6 vowels (*a, e, i, o, u* and syllabic *r*) and 25 consonants (*m, v, n, l, r, j, nj, lj, p, b, f, s, z, c, t, d, ć, đ, š, ž, č, dž, h,* *k*, *g*). The acoustic and articulatory characteristics of the vowels do not change depending on their placement (regardless of whether they are in a short, long, accented or unaccented syllable). In addition to these 6 vowels, there also exist the diphthong *ie*, which is marked in writing as *je* or *ije*.

The prosodic system consists of 4 accents (two long and two short accents, both with a descending and ascending tone) and unaccented post-accentual lengths. The accentual system of the Croatian standard language is neo-štokavian, although it exists today with many differentiations from the prosodic models codified in the second half of the 19th century. Accent location is not fixed to a specific syllable, but the distribution of accents does have some limitations. Croatian has proclitics and enclitics where only proclitics can take over the accent from the accented word with a descending accent in the initial syllable, while enclitics cannot do this. The Croatian standard language is characterised by a number of phonologically (allophones) and morphologically (allomorphs) conditioned alternations leading to a large variation of expressions of morphemes in different word-forms of a single paradigm.

The Croatian standard language differentiates between ten parts of speech, of which five inflect (nouns, adjectives, numbers (partially), pronouns and verbs) and four do not inflect (prepositions, conjunctions, particles and exclamations), while some adverbs inflect only in comparison.

The grammatical categories that characterise the majority of declinable words are gender (three values: masculine, neuter, feminine), number (two values: singular, plural), case (seven values: nominative, genitive, dative, accusative, vocative, locative and instrumental). Some declinable words have special categories that are systematically marked with inflectional endings such as definiteness in adjectives and animacy, but only in accusative singular of masculine nouns and adjectives. Verbs are characterised by the categories of: manner (five values: indicative, imperative, conditional 1, conditional 2, optative), person (three values: 1st, 2nd, 3rd), number (two values: singular, plural), voice (two values: active, passive) and tense (seven values: present, aorist, imperfect, perfect, pluperfect, future 1, future 2). The verbs *biti* ('to be') and *htjeti* ('to will') are auxiliary in Croatian. Verbs also have a complicated aspectual system (imperfective and perfective with additional subvalues such as inchoativity, iterativity, partitivity etc.) and they also encode the feature of transitivity. Adjectives and adverbs can take comparative forms (three values: positive, comparative and superlative). On top of that, there is an extensive internal homography, i.e. on average in nouns 7 cases in singular and 7 cases in plural are represented by 10 different types so there is an overlap in at least 4 cases. Such a rich inflectional system with homography leads to a sparsity of data in Croatian data sets and complicates the NLP approaches. The higher language levels can't be processed without sorting out the inflectional level and this has been done entirely for the first time in (Tadić, 1994).

Words in Croatian are formed by derivation, compounding and rarely conversion. Derivation uses suffix, prefix, and prefix-suffix formation, while compounding uses non-suffix formation, compound suffix formation, coalescence, formation through compound abbreviations. Suffix formation is the most common.

The Croatian language is characterised by an SVO syntactic structure and relatively free word order (permutations of constituents are possible with some limitations, such as clitic placement). It is a basic rule for structuring stylistically unmarked discourse that the first place is taken by the theme (old information), which is followed by the rheme (comment, new information). The subject of a sentence does not have to be explicitly expressed, and its omission is desirable insofar as it is repeated a number of times within a narrow context. Double-negation is required. The agreement of components in gender, number and case is typical of Croatian sentence structure.

Sentence organisation can be both coordinated and subordinated with the usage of appropriate conjunctions or without them. Genitive expressions of possession are avoided in favour of possessive adjectives, and in the modern Croatian the use of preterite tenses (imperfect, aorist and pluperfect) and passive constructions is reduced although imperfect and aorist gained some popularity back, in text messages for their shortness (Žic Fuchs, 2002).

Croatian is the main language used and taught in schools at all levels of education: primary, secondary and higher education with the obligatory state graduation exam after secondary school. The literacy ratio in Croatia is 99.2%.[4]

The Croatian Radio and Television (HRT) is the national television (5 channels) and radio (3 channels at state and 8 channels at regional level) public boradcasting service that broadcasts almost entirely in Croatian with some programmes in minority languages. There are also 7 commercial TV stations with national concession and 20 TV stations with local or regional concession and 36 IPTV/cable/satellite TV stations that all broadcast in Croatian.[5] The foreign movies are always subtitled and only the movies for children and animated movies are synchronised to Croatian. The subtitles of the most popular series have the largest reading public in Croatia, even larger than the largest daily newspapers or the most popular news portals. This fact can certainly impact the importance of multilingual language resources from Croatia.

2.2 Croatian in the Digital Sphere

In Table 1 the usage of ICT in households is presented[6].

Households equipped with ICT	
Personal computer	76
Internet access	82
Type of Internet access in households	
Only fixed broadband	88
Only mobile broadband	66
Broadband total	99

Table 1: Usage of ICT in households and by individuals in 2018 (in %)

The Croatian Web Archive catalogues and stores web resources: portals (news, thematic, etc.), websites of institutions, associations, events, scientific projects, books, journals, etc. from 1998, and today it's run by the National and University Library in collaboration with the Zagreb University Computing Centre (Srce). In 2020, around 90 Tb of data from all web pages under .hr domain was collected.[7]

There is a significant amount of online content found across the websites of state and public bodies, commercial companies etc. as can be observed through web-crawling data collection efforts such as hrWaC v2.1 (Ljubešić and Klubička, 2014), European Language Resource Coordination (ELRC).[8] The first language of web pages is predominantly Croatian with possible translations to English or other, mostly EU languages.

The Croatian Wikipedia (founded in 2003) currently has 209,778 articles (2021-12-19). For the past several years, it has been ranked around 47th place in terms of number of articles.[9]

[4] International Illiteracy Day, Croatian Bureau of Statistics, https://www.dzs.hr/eng/important/Interesting/pismenost.htm (accessed 2021-12-15)

[5] Agency for Electronic Media, https://pmu.e-mediji.hr/Public/PregledTvNakladnici.aspx (accessed 2021-12-15)

[6] Croatian Bureau of Statistics, Usage of ICT in households and by individuals, 2018, The first results, https://www.dzs.hr/Hrv_Eng/publication/2018/02-03-02_01_2018.htm (accessed 2021-12-15)

[7] Croatian Web Archive, https://haw.nsk.hr/en/statistics/ (accessed 2021-12-15)

[8] https://elrc-share.eu, or the Paracrawl project (Bañón et al., 2020)

[9] https://meta.wikimedia.org/wiki/List_of_Wikipedias#All_Wikipedias_ordered_by_number_of_articles (accessed 2021-12-15)

The use of Croatian in social media is growing steadily. Croatian is now prevalently used on platforms such as Facebook, Twitter, LinkedIn, Instagram, YouTube etc. Croatian appears in both Google Translate and Bing Translator as the source and target language, appearing in Google Translate as early as 2008 (Simeon, 2008). In the last three years, with the introduction of NMT methods, the translations are of much higher quality. Most of social media today offer translations of posts in/from Croatian.

Open-source software such as Firefox, Thunderbird, GNULinux, LibreOffice and KDE have all been localised into Croatian by volunteer translators, while Apple, Google and Microsoft offer a localised version of their systems and interfaces for all their software services (MacOS, iOS, Google Documents etc., Microsoft Windows and Office etc.).

3 What is Language Technology?

Natural language[10] 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 specialized 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 approached. We are gradually moving from a methodology in which a pipeline of multiple

[10] 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).

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 realizing 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.[11]

[11] 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 Croatian

Although the Croatian missed the boost in the first wave of LT development in the 1990s due to the defence in the Croatian Homeland War (1991-1995) and consequently exclusion from all EU funding, the Croatian LT community still tried to catch up with other European languages in the 2000s thanks to some national funding. That development was described in (Tadić et al., 2012). In the past ten years the support for the development of Croatian language technology advanced primarily because of Croatia joining the EU in 2013. For few years before and after this event, a focus was set on the rapid development of LTs for Croatian since it was lagging behind from other official EU languages that joined earlier. The position of 24th official EU language resulted in regular inclusion of Croatian into large multilingual NLP campaigns and shared tasks at different NLP levels and it started to be performed by non-Croatian NLP experts also. Although in some areas there are still a number of fundamental language resources not yet available for Croatian, the progress has been made in the area of language resources collection, text analytics, language models, computer assisted language learning (CALL) and machine translation (MT), while the speech processing is still seriously underdeveloped. A number of short term projects funded by EU and national funds were situated mostly in academic institutions, while joining the CLARIN ERIC was also an event that contributed to further development. Fundamental building blocks such as lemmatisation, MSD tagging, NERC and syntactic analysis tools have been provided, but in terms of real-life usefulness, for some tasks the training datasets are still too small to build robust and reliable industrial strength systems. From a natural language understanding perspective, there is a new version of Croatian Wordnet (v2.1) (Šojat et al., 2018a) and in 2016, a layer of semantic roles was added to the Croatian Dependency Treebank[12] thus providing the basic LRs for further semantic processing at lexical and clausal levels. The following sections summarise the corpora, tools and services that have been developed for Croatian in last ten years and that can usually be found in different LT repositories such as META-SHARE, CLARIN, ELRC-SHARE etc. The versions of language resources cited are the newest versions.

4.1 Language Data

Monolingual Corpora

After the Croatian National Corpus v3[13] (Tadić, 2009) was released in 2013, significant advances in large corpora collection have been made when a number of such corpora were compiled. They were primarily aimed as the necessary data sets for different NLP purposes, such as general use hrWac v2.1 (Ljubešić and Klubička, 2014), for word embeddings (Shekhar et al., 2020), for loanwords detection (Bogunović et al., 2021), ParlaMint-HR 2.1 for investigations of the parliamentary genre (Erjavec et al., 2021), MARCELL Croatian legislative subcorpus (Váradi et al., 2020) and other.

Also, a number of smaller specialised corpora with particular uses were compiled such as social media research, e. g. tweet processing (Ljubešić et al., 2019), for training the basic tools (Ljubešić et al., 2018), for sentiment analysis (Thakkar et al., 2021a), for investigation of speakers with disorders (Kuvač Kraljević et al., 2020) or learner corpora[14] (Mikelić Preradović et al., 2015).

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).

[12] https://hobs.ffzg.hr

[13] http://filip.ffzg.hr

[14] http://teitok.clul.ul.pt/croltec/

Bilingual and Multilingual Corpora

The bilingual or parallel corpora where Croatian was one of the languages in a pair were either produced as stand-alone parallel/comparable corpora or they represented results of wider campaigns to collect parallel data. The examples for the former are hrenWaC 2.0 (Ljubešić et al., 2016a), bi-texts in turistic domain (Toral et al., 2016) or MARCELL Croatian-English Parallel Corpus of Legislative Texts (Váradi et al., 2020). The examples of the latter are ParaCrawl9, Bible translations[15] or texts related to COVID-19 pandemic, then a number of parallel corpora and TMs collected from different Croatian public institutions where documents and/or web pages were treated as open data sources, following the PSI directive. These were mostly crawled or handed over during the ELRC data collection campaigns, so they could be found easily in ELRC-SHARE. Significant contribution in terms of numbers of TUs represent bilingual parallel corpora compiled for training of NMT models within the NTEU project.[16]

After July 1st, 2013 and joining the EU, Croatian regularly became a language of interest in large multilingual data collection campaigns and shared tasks. Here we would mention just a few: Universal Dependencies (UD)[17] (Zeman and al, 2021), C4Corpus (with 7 different versions depending on licenses) (Gurevych et al., 2016), Deltacorpus (Mareček et al., 2016), EU Patents translations, EU EAC TM,[18], JRC DGT TMs (Ljubešić and Erjavec, 2018) ParlaMint comparable corpora (Erjavec et al., 2021) and Comparable Wikipedias of South Slavic Languages (Ljubešić et al., 2021), OSCAR,[19] SETimes,[20] TED talks,[21] OPUS,[22] W2C (Majliš, 2011), WikiMatrix.[23]

Lexical Resources

The largest freely available lexical resources are inflectional lexicons: *Croatian Morphological Lexicon* (HML)[24] (Tadić, 2005) and *hrLEX v1.3* (Ljubešić et al., 2016b). Unfortunately, there is only one general language dictionary freely available for online search in its fullest extent: *Hrvatski jezični portal* (*Croatian Language Portal*)[25] with the lexicographical base of a commercial lexicographical publisher Znanje behind it. Other general language dictionaries are either only partially accessible, e. g. *Mrežnik* (*Croatian Net Dictionary*),[26] or are packed into a proprietary app for mobile or non-mobile devices like Rječnici[27] by the commercial publisher Školska knjiga. The *Školski rječnik (School Dictionary)*[28] is compiled for the school children in order to serve as the prescriptive dictionary.

Other larger lexica are specialised like the Croatian Old Dictionaries Portal,[29] or Dictionary

[15] https://opus.nlpl.eu/bible-uedin.php
[16] https://nteu.eu
[17] https://universaldependencies.org
[18] https://huggingface.co/datasets/europa_eac_tm
[19] https://oscar-corpus.com
[20] http://nlp.ffzg.hr/resources/corpora/setimes/
[21] https://wit3.fbk.eu/home
[22] https://opus.nlpl.eu
[23] https://opus.nlpl.eu/WikiMatrix.php
[24] https://hml.ffzg.hr and HML v5.0 at http://metashare.ilsp.gr:8080/repository/browse/croatian-morphological-lexicon-v50/2d429672703d11e28a985ef2e4e6c59e27b37c59b92d42a5be839f7daff7ecfb/
[25] https://hjp.srce.hr
[26] http://ihjj.hr/mreznik/
[27] https://www.rjecnici.hr
[28] http://rjecnik.hr
[29] http://crodip.ffzg.hr/default_e.aspx

of Neologisms,[30] Spelling dictionary,[31] Dictionary of Phrasemes,[32] the Valency Dictionary,[33] the Collocation Dictionary[34] and the Croatian Terminology Portal,[35] which offers central access to various terminological dictionaries and a list of other terminology resources.[36] A specific type of lexical resource is the Croatian Derivative Lexicon – CroDeriv[37] (Šojat et al., 2012, 2013, 2014; Filko et al., 2020) and DerivBase.HR (Šnajder, 2014) that represent the first steps of processing at the level of derivative morphology. Both have been connected later with the Universal Derivations initiative[38] but the main difference is that in CroDeriv each entry is manually checked while the DerivBase.HR has been compiled using unsupervised and language-model-filtered machine learning approaches.

Grammars and Models

After releasing NooJ[39] as a freely accessible multi-platform framework for developing and for the application of formal grammars (Silberztein et al., 2012) in 2012, the development of NooJ grammar models accelerated (Srebačić et al., 2015; Bekavac et al., 2015; Šojat et al., 2016, 2018b; Karl et al., 2018; Landsman Vinković and Kocijan, 2020; Thakkar et al., 2021b), also because it was included more in the teaching at under- and graduate-level of studies of linguistics and information sciences at the University of Zagreb, Faculty of Humanities and Social Sciences.

Following the initiatives to collect and annotate LRs at different language levels using the universal and common annotation framework (e. g. Universal Tagset, Universal Dependencies or Universal Derivation), in 2020 it was suggested by (Alves et al., 2020) to establish a similar Universal NER framework called UNER.

In last couple of years, after the heavier introduction of language models (LM) approaches in NLP initiated mostly by NMT paradigm and by BERT[40] (Devlin et al., 2019), a similar model was built for Croatian but usually in combination with other languages, such as CroSloEngualBERT (Ulčar and Robnik-Šikonja, 2020a,b); BERTić (Ljubešić and Lauc, 2021) or ELMo embeddings models (Ulčar, 2019). In CLEOPATRA MSC project a set of LMs for Croatian is being developed while the sentiment analysis for EU official Slavic languages will be tackled with the usage of large LMs as well (Thakkar and Pinnis, 2020).

4.2 Language Technologies and Tools

Tools

A number of tools and services are available for the Croatian language already, but somehow they didn't find its way entirely to the most popular NLP suites, platforms or pipelines such as spaCy, FreeLing, NLP Cube, TextRazor, Cloud Natural Language, Apache Open NLP, or just partially, like in Lexalytic platform, Stanford-NLP, etc. However, there is the whole Croatian pipeline[41] developed within the UD initiative, namely UDPipe, and it found its way also into

[30] http://rjecnik.neologizam.ffzg.unizg.hr
[31] https://pravopis.hr
[32] http://frazemi.ihjj.hr
[33] http://valencije.ihjj.hr
[34] http://ihjj.hr/kolokacije/
[35] http://nazivlje.hr
[36] http://nazivlje.hr/english/page/other-terminology-sources/9/
[37] http://croderiv.ffzg.hr
[38] https://ufal.mff.cuni.cz/universal-derivations
[39] http://nooj4nlp.org
[40] https://github.com/google-research/bert
[41] https://ufal.mff.cuni.cz/udpipe

the GATE platform as a POS-tagger for Croatian,[42] as well as Weblicht[43] platform. The UD data served also to produce the Croatian segment in the UDify[44] (Kondratyuk and Straka, 2019), a multilingual LM for morphological and syntactical processing also providing the UD types of data (UPOS, UFeats, Lemmas, Deps).

Apart from the Croatian Language Processing Pipeline[45] developed back in 2013 during the CESAR project as a part of the META-NET initiative and available through META-SHARE, the CLASSLA fork for Stanford Stanza pipeline[46] for processing South Slavic languages has been developed as well.

Also, at the lexical and event semantics level, two popular online services feature, among other languages, also process Croatian texts, namely, Wikifier[47] and Event Registry.[48] In Babelnet [49] Croatian is well represented and it is ranked 41th with 2,933,659 synsets.

Translation Technologies

Support for Croatian as a source and target language in MT systems was provided as early as in MT@EC, followed by CEF AT and later within eTranslation services. This development was enabled by initial collection of resources either from monolingual (projects MARCELL and CURLICAT) sources, comparable (project ACCURAT) or parallel corpora (projects Let'sMT! and ABU-MATRAN as well as DGT TMs). The introduction of NMT paradigm upgraded the quality of translations and this can be seen particularly with the results of the CEF project *EU Council Presidency Translator*[50] that was developed for the Croatian EU Presidency in the first half of 2020. This service was the first one by that consortium, that fully implemented NMT method and it became very popular in Croatia and elsewhere for its high quality translations. The system had beaten Google Translate in hr → en and en → hr directions for several BLEU points and in the first year of its usage it translated more than 60 million tokens.

Computer Assisted Language Learning (CALL)

From 2015 to 2016 within the ESF-funded project HR4EU the *HR4EU, a Portal for Learning Croatian as a Foreign Language*[51] was produced (Filko et al., 2016; Farkaš et al., 2016). The portal provides a free fully equipped place for teaching yourself Croatian with three courses (beginners, intermediate and advanced). Each of the courses is composed of lectures, quizzes, questionnaires and exams. The lectures are accompanied by more than 200 illustrations that are combined with the text. A set of important Croatian LRs is seamlessly included in this portal, so the learners are directed towards the extensive usage of Croatian National Corpus, Croatian Wordnet, Croatian Depencency Treebank. The users are also educated through a series of short explanatory video clips. Although already more than five years passed after the completion of the project, the portal is still functioning with more than 12,100 registered users so far from all around the world.

From 2018 the Centre for Croatian as a Foreign Language (CROATICUM) of the Faculty of Humanities and Social Sciences, University of Zagreb, in collaboration with the Central State Office for Croats Abroad, produced two online courses for Croatian Language at levels A1[52]

[42] https://cloud.gate.ac.uk/shopfront/displayItem/tagger-pos-hr-maxent1

[43] https://weblicht.sfs.uni-tuebingen.de/weblicht/

[44] https://github.com/Hyperparticle/udify

[45] https://lt.ffzg.hr

[46] https://github.com/clarinsi/classla

[47] https://wikifier.org

[48] https://eventregistry.org

[49] https://babelnet.org

[50] https://hr.presidency.eu

[51] https://hr4eu.eu

[52] https://a1.ffzg.unizg.hr

and A2.[53] This two free online courses serve as the supporting material for the official Croatian language live courses held at the CROATICUM, but on the commercial basis.

Speech Processing

Speech technology is the most underdeveloped area for Croatian, although there were some attempts at the Department of Phonetics, Faculty of Humanites and Social Sciences, University of Zagreb and at the Faculty of Electrical Engineering and Computing of the same university, no demo systems for AST or TTS exist, let alone the industrial strength applications. The phoneme model was produced back in 1998 in the EU-funded project MBROLA, but after that nothing much happened. For this reason commercial players (e. g. Dragon Systems, Newton Technologies, Alfanum, etc.) started to offer speech modules for Croatian. The support in Android-based mobile devices is provided at the level of the operating system, but it does not exist in iOS environment, i. e. Siri still does not use Croatian. Even the multimodal resources are scarce and rarely produced by local experts. Apart from the Croatian Weather Dialogue Corpus published in 2013 (Načinović et al., 2009), the Collins Multilingual databases (MLD) – WordBank and PhraseBank with audio files for 28 languages (incl. Croatian) have been compiled in 2016, while the GlobalPhone Croatian Pronunciation Dictionary has been compiled in 2013. The Croatian data in TalkBank[54] are closing this rather limited set of speech data for Croatian.

4.3 Projects, Initiatives, Stakeholders

In Croatia there used to be a nationally funded programme for LT from 2007 to 2012 (Dalbelo Bašić et al., 2007) which set foundations for widening of the research from the Faculty of Humanities and Social Sciences, University of Zagreb to a number of other public institutions in Croatia that became relevant in LT, such as the Faculty of Electrical Engineering and Computing, University of Zagreb, Institute of Croatian Language and Linguistics, University or Rijeka, University of Split, University of Zagreb Computing Centre (SRCE), but it also involved private companies, such as Ciklopea or Integra. Croatian Language Technologies Society, established 2004, has a mission to loosely coordinate LT activities in Croatia.

Unfortunately, the draft of the national strategy for AI was proposed in late 2019, but it was withdrawn after announcement of changed EU regulations and new directive in 2021. Now a new strategy is expected to be drafted taking into account changed EU regulations.

However, the dominant role regarding the further development of Croatian LT in previous decade was played by the EU through its FP7, ICT-PSP, H2020 and CEF programmes funding involvement of several Croatian research teams. A number of projects resulted in increased research and development activity in the field of LT, that predominantly still remains in the academic circles and rarely involves industry. To list just a few of the projects that were relevant: CLARIN[55] (FP7 RI), ACCURAT[56] (FP7), Lets'MT![57] (ICT-PSP), XLike[58] (FP7), ABUMATRAN[59] (FP7), HR4EU[60] (ESF), MARCELL[61] (CEF), CLEOPATRA[62] (H2020 MSC), EU Council

[53] https://a2.ffzg.unizg.hr
[54] https://ca.talkbank.org/access/Croatian.html
[55] https://clarin.eu
[56] http://accurat-project.eu
[57] http://www.letsmt.org
[58] https://xlike.ijs.si
[59] https://abumatran.eu
[60] https://www.hr4eu.hr
[61] https://marcell-project.eu
[62] https://cleopatra-project.eu

Presidency Translator[63] (CEF), CURLICAT[64] (CEF), NLTP (CEF), NEC-TM[65] (CEF), PRINCIPLE[66] (CEF) and several bilateral and COST-actions. Joining CLARIN ERIC as a full member in 2018 provided additional kick in activities, particularly in the view of LT serving as Research Infrastructure in Humanities and Social Sciences.

At the national level large projects were funded through the Croatian Research Council such as umbrella project Struna[67] where terminological collections from many domains are compiled, Repository of metaphors,[68] Collocations Dictionary, Mrežnik (Croatian Net Dictionary), etc.

A Remark about Serbo-Croatian Language Resources

There is no state in the world that in its constitution lists Serbo-Croatian as an official language – the official language in Serbia is Serbian, in Montenegro Montenegrin, in Croatia Croatian and in Bosnia and Herzegovina three languages have the official status: Bosnian, Croatian and Serbian. However, for several language resources, that are in use in the NLP community, the Serbo-Croatian language code is used to denote the so called macro-language or simply to legitimate the deliberate mixture of the four mentioned, separated, standard languages. The existence of e. g. Serbo-Croatian Wikipedia contributes largely to this set of language resources where its dumps are sometimes used for corpus collection or training of different language models. The quality of such resources will always be dubious since it is expected to render poorer results, much like as if someone would try today to build e. g. a corpus for Czechoslovakian or even Hindustani while the clear cut between Czech and Slovak as well as Hindi and Urdu has been already well established and recognised. Unfortunately, the usage of such compound names has often been used for achieving some political goals or simply as a convenience stemming out of colonial attitude (e. g. Czechoslovakian was forced between two world wars supporting the unification of Czechoslovakia; similarly, the official name of the language in the Kingdom of Yugoslavia in the same time was Serbo-Croato-Slovenian, whilst in the communist Yugoslavia after the WWII it turned into Serbo-Croatian; the term Hindustani was predominantly used during the times of unified British colonial control over the area that today encompasses three states – India, Pakistan and Bangladesh – and two major religions – Hinduism and Islam). In this respect, the composed language names not just introduce confusion, but often denote the artificially or theoretically constructed "languages" or remnants of long passed linguistic situations. These names should be avoided in language technologies since the language resources should be compiled from empirically attested real language data and not hypothetically derived constructs. Such virtual convergence of individual language resources may in the long run result in decreasing of the language variety and this is certainly not what the ELE project would like to achieve, quite the contrary. The very existence of such language resources (with compound names), does not automatically mean that they could be used uncritically, with no questions asked, e. g. "Why parallel language labels – Serbian and Serbo-Croatian – exist at all?". Believing that Serbo-Croatian language resources would contribute to the quality of derived results when processing individual Bosnian, Croatian, Montenegrin or Serbian language data, is more than an optimistic view.[69] Therefore, we would strongly recommend

[63] https://hr.presidencymt.eu

[64] https://curlicat.eu

[65] https://www.nec-tm.eu

[66] https://principleproject.eu

[67] http://struna.ihjj.hr

[68] http://ihjj.hr/metafore/

[69] This position does not prevent the building of, e. g. multilingual language models where data from many languages is combined, but they should use clearly separated data from individual languages and not the mixture of language data where the language features and language identity is obliterated for the sake of maintaining

to all LT researchers to avoid the usage of Serbo-Croatian language resources for any kind of processing of the Croatian language data, since the results will surely be noisy and much more error-prone than using only language resources labelled with the Croatian language code alone.

5 Cross-Language Comparison

The LT field[70] 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[71] 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.)

the constructed macro-language, sometimes even for non-scientific reasons.

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

[71] 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[72]

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[73] 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.[74]

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

[72] 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.

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

[74] 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 2 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g. German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[75] 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 2 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

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

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

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

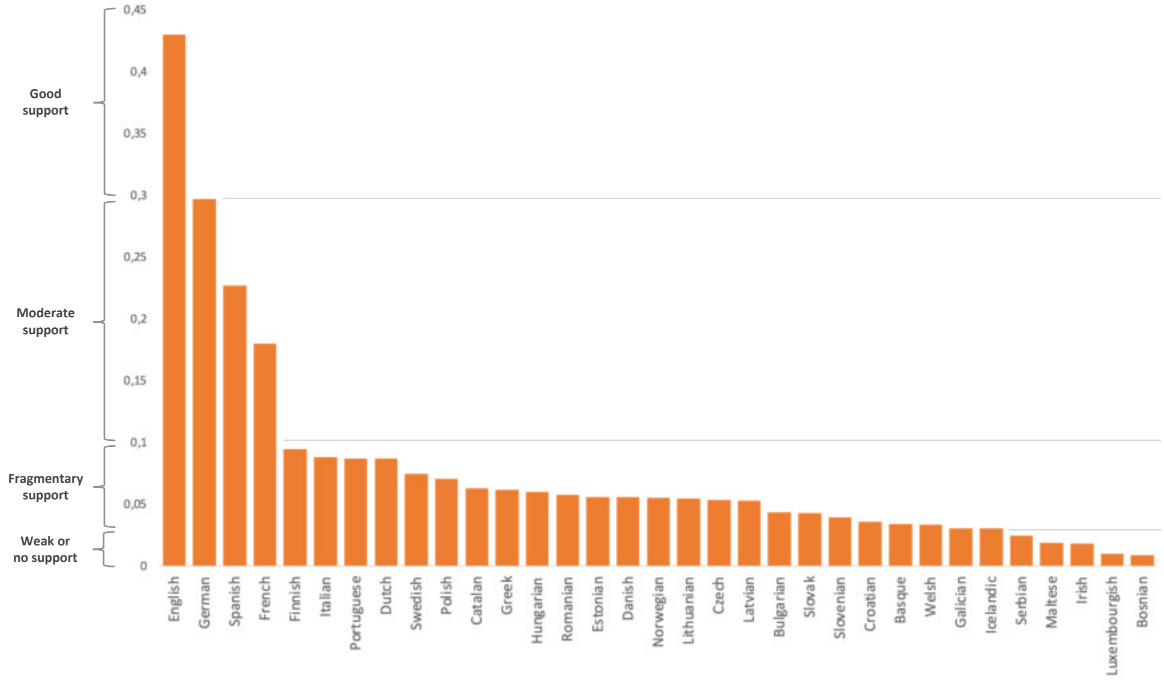

Figure 2: Overall state of technology support for selected European languages (2022)

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

Technological support for Croatian has progressed in a number of areas of LT in the past decade compared to the state of affairs described in the META-NET White Paper (Tadić et al., 2012). Digital language resources have both increased in number and volume while they also improved in quality and variety. Resources, basic NLP tools and LT services are provided by academia, research institutes and occasionally private companies as outputs of various research projects, usually coordinated by academic institutions, predominantly funded by EU or national funds, and rarely self-funded (e. g. University of Zagreb internal single-year projects). Some significant progress has been made with respect to available corpora and lexica, language models, text processing tools, and machine translation, while there is still a serious underdevelopment in the subfield of speech processing (both synthesis and recognition). The available datasets origin from a variety of sources and they cover several thematic domains, text types; they are available as raw or annotated, and come as monolingual, bi- or multilingual resources. However, their individual size is lagging behind in terms of appropriateness for building large language models or robust, ready to use tools and applications.

Untapped Potential and Open-Source Culture

There is much untapped and currently inaccessible data that could make a huge impact on the future of Croatian LT, if collected and applied appropriately. For example, there are a lot of textual data produced by the different public authorities that are still not recognised as valuable language data. The series of ELRC workshops which focused on the PSI Directive and its applications, greatly contributed to the enhanced collaboration of LT researchers with different state, regional and local administrations. Another source of valuable language and speech data, that has not been addressed so far, is aligned audio and subtitle text data stored in the archives of the national broadcaster (Croatian Radio and Television), that could be used to build multimodal processing systems. Additionally, the Croatian Scientific Journals Portal[76] also provides a valuable source of documents (more than 250,000 articles from more than 500 journals with estimated 650 million tokens) that are published under a variety of permissive licenses, i. e. different levels of open access. It is being processed right now, while this report is being written.

Long-term Strategy

Although AI is already a part of our everyday lives – when we use language technologies for browsing the internet, shopping online, interacting with smart devices and appliances, etc. – we still lack out-of-the-box general-oriented systems to communicate digitally in Croatian. There is no doubt that there have been ample developments in LT over the past ten years, but, as described in this report, many commonly used and necessary technologies are still not available for Croatian (general speech processing, human-computer interaction, natural language understanding, multimodal processing, etc.) and, if some advance in technologies is recorded, there are no available applications (summarisation, question answering, etc.). Many technologies are more advanced abroad and Croatian became a part of some multilingual systems for semantic analysis, machine translation, and speech processing, but in other cases Croatian was not included in such multilingual systems for different reasons.

The national strategy for AI was drafted in 2019, but it was withdrawn after announcement of change in EU regulations and new directive in 2021. Now, a new strategy is expected to be drafted taking into account the changed EU regulations.

One of the long-term plans is to secure the presence of Croatian NLP modules in the major NLP platforms (commercial and non-commercial) such as spaCy, FreeLing, NLP Cube, TextRazor, Cloud Natural Language, Apache Open NLP, etc., in order to secure the sustainability and wider usage of LT for Croatian and, consequently, its digital equality with other languages.

Collaboration

Croatian Language Technologies Society was established in 2004 and has a mission to loosely coordinate the advancements of LT development in Croatia, but also abroad (e. g. it is a Croatian partner in ELRC initiative). Participation by Croatian LT research teams and industrial partners in a number of EU-funded projects enabled the widening of the collaboration from the national to the international level. This helped a lot in collecting experience from similar research teams that were in advanced stages of the development of LT for their languages and also enabled easier tackling of similar problems in the development of LT for Croatian. Also, participation and collaboration in pan-European research infrastructures and initiatives, e. g. CLARIN ERIC or ELRC, gave a necessary push in the back to the Croatian LT community.

[76] https://hrcak.srce.hr

Vision

Although a number of technologies and resources for Croatian already exist, there are expectedly less technologies and resources for the Croatian language than for English and some other European languages, such as German, French, Italian, Spanish. However, it is to be expected that LTs for Croatian language and speech will be developed at least to the level that will allow Croatian to be in the same category with the EU official languages of comparable number of speakers by 2030. This Report on the Croatian Language surely presents an important step in that direction.

Acknowledgments

The author's gratitude goes to Matea Filko, Tea Vojtěchová, Annika Grützner-Zahn and Maria Giagkou. This report has benefited from their insightful comments.

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.

Diego Alves, Tin Kuculo, Gabriel Amaral, Gaurish Thakkar, and Marko Tadić. Uner: Universal named-entity recognition framework. In *Proceedings of the 1nd International Workshop on Cross-lingual Event-centric Open Analytics*, pages 72–79, Heraklion, Greece, June 2020. CEUR Workshop Proceedings. URL http://ceur-ws.org/Vol-2611/short1.pdf.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417.

Božo Bekavac, Kristina Kocijan, and Marko Tadić. Near Language Identification Using NooJ. In Johanna Monti, Max Silberztein, Mario Monteleone, and Maria Pia di Buono, editors, *Formalising Natural Languages with NooJ 2014: Selected papers from the NooJ 2014 International Conference*, pages 152–166, Sassari, Italy, 2015. Cambridge Scholars Publishing, Newcastle upon Tyne, UK.

Irena Bogunović, Mario Kučić, Nikola Ljubešić, and Tomaž Erjavec. Corpus of Croatian news portals ENGRI (2014-2018). In *Slovenian language resource repository CLARIN.SI*. 2021. URL http://hdl.handle.net/11356/1416.

2011 Census. *Statistical Reports 1469*. Croatian Bureau of Statistics, Zagreb, 2013. URL https://www.dzs.hr/Hrv_Eng/publication/2012/SI-1469.pdf. (accessed 2021-12-15).

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

Bojana Dalbelo Bašić, Zdravko Dovedan, Ida Raffaelli, Sanja Seljan, and Marko Tadić. Computational linguistic models and language technologies for croatian. In Vesna Lužar-Stiffler and Vesna Hljuz Dobrić, editors, *Proceedings of the 29th International Conference on Information Technology Interfaces (ITI 2007)*, pages 521–528, Cavtat, Croatia, June 2007. Srce, Zagreb.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darģis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Roberto Bartolini, Andrea Cimino, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. In *Slovenian language resource repository CLARIN.SI*. 2021. URL http://hdl.handle.net/11356/1431.

Daša Farkaš, Matea Filko, and Marko Tadić. Hr4eu – using language resources in computer aided language learning. In *Proceedings of the Second International Conference Computational Linguistics in Bulgaria*, pages 38–44, Sofia, Bulgaria, September 2016. The Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences.

Matea Filko, Daša Farkaš, and Diana Hriberski. Hr4eu – a web-portal for e-learning of croatian. In Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny, editors, *CALL communities and culture – short papers from EUROCALL 2016*, pages 137–143, Limassol, Cyprus, August 2016. Research-publishing.net, Ulster. URL https://files.eric.ed.gov/fulltext/ED572005.pdf.

Matea Filko, Krešimir Šojat, and Vanja Štefanec. The Design of Croderiv 2.0. *The Prague Bulletin of Mathematical Linguistics*, (115):83—104, 2020. doi: 10.14712/00326585.006.

Iryna Gurevych, Ivan Habernal, and Omnia Zayed. C4Corpus. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*. 2016. URL http://hdl.handle.net/11372/LRT-2209.

Dario Karl, Božo Bekavac, and Ida Raffaelli. A construction grammar approach in the nooj framework: Semantic analysis of lexemes describing emotions in croatian language. In Ignazio Mauro Mirto, Mario Monteleone, and Max Silberztein, editors, *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*, pages 114–123, Palermo, Italy, 2018. Springer. doi: 10.1007/978-3-030-10868-7_11.

Dan Kondratyuk and Milan Straka. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1279.

Jelena Kuvač Kraljević, Gordana Hržica, and Lana Kologranić Belić. Croatian Corpus of Non-Professional Written Language – Typical speakers and speakers with language disorders. *Govor*, 37(2):125—147, 2020.

Mirela Landsman Vinković and Kristina Kocijan. Preparing the nooj german module for the analysis of a learner spoken corpus. In Božo Bekavac, Kristina Kocijan, Max Silberztein, and Krešimir Šojat, editors, *Formalising natural languages: applications to natural language processing and digital humanities. NooJ 2020*, pages 146–158, Zagreb, Croatia, 2020. Springer. doi: 10.1007/978-3-030-70629-6_13.

Nikola Ljubešić and Tomaž Erjavec. JRC EU DGT Translation Memory Parsebank DGT-UD 1.0. In *Slovenian language resource repository CLARIN.SI*. 2018. URL http://hdl.handle.net/11356/1197.

Nikola Ljubešić and Filip Klubička. bs,hr,srWaC - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-0405. URL https://aclanthology.org/W14-0405.

Nikola Ljubešić and Davor Lauc. BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Bogdan Babych, Olga Kanishcheva, Preslav Nakov, Jakub Piskorski, Lidia Pivovarova, Vasyl Starko, Josef Steinberger, Roman Yangarber, Michał Marcińczuk, Senja Pollak, Pavel Přibáň, and Marko Robnik-Šikonja, editors, *Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine, April 2021. Association for Computational Linguistics (ACL).

Nikola Ljubešić, Miquel Esplà-Gomis, Sergio Ortiz Rojas, Filip Klubička, and Antonio Toral. Croatian-English parallel corpus hrenWaC 2.0. In *Slovenian language resource repository CLARIN.SI*. 2016a. URL http://hdl.handle.net/11356/1058.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4264–4270. European Language Resources Association (ELRA), May 2016b.

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. Training corpus hr500k 1.0. In *Slovenian language resource repository CLARIN.SI*. 2018. URL http://hdl.handle.net/11356/1183.

Nikola Ljubešić, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1. In *Slovenian language resource repository CLARIN.SI*. 2019. URL http://hdl.handle.net/11356/1241.

Nikola Ljubešić, Filip Markoski, Elena Markoska, and Tomaž Erjavec. Comparable corpora of South-Slavic Wikipedias CLASSLA-Wikipedia 1.0. In *Slovenian language resource repository CLARIN.SI*. 2021. URL http://hdl.handle.net/11356/1427.

Martin Majliš. W2C – Web to Corpus – Corpora. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*. 2011. URL http://hdl.handle.net/11858/00-097C-0000-0022-6133-9.

David Mareček, Zhiwei Yu, Daniel Zeman, and Zdeněk Žabokrtský. Deltacorpus 1.1. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*. 2016. URL http://hdl.handle.net/11234/1-1743.

Nives Mikelić Preradović, Monika Berać, and Damir Boras. Learner corpus of croatian as a second and foreign language. In Kristina Cergol Kovačević and Sanda Lucija Udier, editors, *Multidisciplinary Approaches to Multilingualism*, pages 107–126, Frankfurt am Main, Germany, 2015. Peter Lang.

Lucia Načinović, Sanda Martinčić-Ipšić, and Ivo Ipšić. Statistical language models for croatian weather-domain corpus. In Hrvoje Stančić, Sanja Seljan, David Bawden, Jadranka Lasić-Lazić, and Aida Slavić, editors, *INFuture2009: Digital Resources and Knowledge Sharing - Proceedings*, pages 333–340, Zagreb, Croatia, November 2009. University of Zagreb, Faculty of Humanities and Social Sciences.

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.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1):49—79, 2020.

Max Silberztein, Tamás Váradi, and Marko Tadić. Open source multi-platform nooj for nlp. In Christian Kay, Martin ; Boitet, editor, *Proceedings of COLING2012*, pages 401–408, Mumbai, India, dec 2012. The COLING2012 Organizing Committee.

Ivana Simeon. *Vrednovanje strojnoga prevođenja (Evaluation of the Machine Translation)*. PhD Thesis. Sveučilište u Zagrebu, Filozofski fakultet, Zagreb, 2008.

Matea Srebačić, Krešimir Šojat, and Božo Bekavac. Croatian derivational patterns in nooj. In Johanna Monti, Max Silberztein, Mario Monteleone, and Maria Pia di Buono, editors, *Formalising Natural Languages with NooJ 2014: Selected papers from the NooJ 2014 International Conference*, pages 55–62, Sassari, Italy, 2015. Cambridge Scholars Publishing, Newcastle upon Tyne, UK.

Marko Tadić. *Računalna obrada morfologije hrvatskoga književnoga jezika (Computational Processing of the Morphology of the Croatian Standard Language)*. PhD Thesis. Sveučilište u Zagrebu, Filozofski fakultet, Zagreb, 1994.

Marko Tadić. The Croatian Lemmatization Server. *Southern Journal of Linguistics*, 29(1/2):206–217, 2005. URL https://www.researchgate.net/publication/228621994_The_Croatian_lemmatization_server.

Marko Tadić. New version of the croatian national corpus. In Dana Hlaváčková, Aleš Horák, Klara Osolsobě, and Pavel Rychlý, editors, *After Half a Century of Slavonic Natural Language Processing*, pages 199–205, Brno, Czechia, 2009. Masaryk University.

Marko Tadić, Dunja Brozović-Rončević, and Amir Kapetanović. *Hrvatski Jezik u Digitalnom Dobu – The Croatian 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/croatian. Georg Rehm and Hans Uszkoreit (series editors).

Gaurish Thakkar and Mārcis Pinnis. Pretraining and fine-tuning strategies for sentiment analysis of latvian tweets. In Andrius Utka, Jurgita Vaičenonienė, Jolanta Kovalevskaitė, and Danguolė Kalinauskaitė, editors, *Human Language Technologies – The Baltic Perspective*, pages 55–61, Kaunas, Lithuania, 2020. IOS Press. doi: 10.3233/FAIA200602. URL https://ebooks.iospress.nl/volumearticle/55523.

Gaurish Thakkar, Nives Mikelić Preradović, and Marko Tadić. Multi-task learning for cross-lingual sentiment analysis. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics*, pages 76–84, Ljubljana, Slovenia, April 2021a. CEUR Workshop Proceedings. URL http://ceur-ws.org/Vol-2829/short1.pdf.

Gaurish Thakkar, Nives Mikelić Preradović, and Marko Tadić. Negation detection using nooj. In Karolj Skala, editor, *Proceedings of the 44th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2021*, pages 263–267, Rijeka, Croatia, 2021b. Croatian Society for Information, Communication and Electronic Technology – MIPRO. URL http://www.mipro.hr/LinkClick.aspx?fileticket=RAMnkK6T5UQ%3d&tabid=196&language=hr-HR.

Antonio Toral, Miquel Esplà-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino, and Andy Way. Tourism English-Croatian Parallel Corpus 2.0. In *Slovenian language resource repository CLARIN.SI*. 2016. URL http://hdl.handle.net/11356/1049.

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.

Matej Ulčar. ELMo embeddings models for seven languages. In *Slovenian language resource repository CLARIN.SI*. 2019. URL http://hdl.handle.net/11356/1277.

Matej Ulčar and Marko Robnik-Šikonja. CroSloEngual BERT 1.1. In *Slovenian language resource repository CLARIN.SI*. 2020a. URL http://hdl.handle.net/11356/1330.

Matej Ulčar and Marko Robnik-Šikonja. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Slovenian language resource repository CLARIN.SI*. 2020b.

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 Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3761–3768, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.464.

Daniel Zeman and al. Universal Dependencies 2.8.1,. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*. 2021. URL http://hdl.handle.net/11234/1-3687.

Jan Šnajder. Derivbase.hr: A high-coverage derivational morphology resource for croatian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3371–3337, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

Krešimir Šojat, Matea Srebačić, and Marko Tadić. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, 00(1):111–142, 2012.

Krešimir Šojat, Matea Srebačić, and Vanja Štefanec. CroDeriV i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika*, 39(75):75–96, 2013.

Krešimir Šojat, Matea Srebačić, Marko Tadić, and Tin Pavelić. Croderiv: a new resource for processing croatian morphology. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3366––3370, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

Krešimir Šojat, Božo Bekavac, and Kristina Kocijan. Detection of verb frames with nooj. In Barone Linda, Mario Monteleone, and Max Silberztein, editors, *Automatic processing of natural-language electronic texts with NooJ : revised selected papers*, pages 157–168, České Budějovice, Czechia, 2016. Springer.

Krešimir Šojat, Matea Filko, and Antoni Oliver. Further expansion of the croatian wordnet. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, pages 356–361, Singapore, Singapore, January 2018a. The Global WordNet Association. URL http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/gwc-2018-proceedings.pdf.

Krešimir Šojat, Kristina Kocijan, and Božo Bekavac. Identification of croatian light verb constructions with nooj. In Samir Mbarki, Mohammed Mourchid, and Max Silberztein, editors, *Formalizing natural languages with NooJ and its natural language processing applications : revised selected papers*, pages 96–107, Rabat, Morocco, 2018b. Springer. doi: 10.1007/978-3-319-73420-0_8.

Milena Žic Fuchs. Communication Technologies and their Influence on Language: An Example from Croatian. *Studia Romanica et Anglica Zagrabiensia*, 47-48:597–608, 2002. URL https://hrcak.srce.hr/file/33157.