EUROPEAN LANGUAGE EQUALITY

D1.27

Report on the Polish Language

Authors	Maciej Ogrodniczuk, Piotr Pęzik, Marek Łaziński, Marcin Miłkowski
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	D1.27
Deliverable title	Report on the Polish Language
Type	Report
Number of pages	24
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Maciej Ogrodniczuk, Piotr Pęzik, Marek Łaziński, Marcin Miłkowski
Reviewers	Jaroslava Hlavacova, Teresa Lynn
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Institute of Computer Science, 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
AGH	Akademia Górniczo-Hutnicza (AGH University of Science and Technology)
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DARIAH	Digital Research Infrastructure for the Arts and Humanities
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019–2022)
ELRC	European Language Resource Coordination (EU project, 2015–2022)
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IPI PAN	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Institute of Computer Science, Polish Academy of Sciences)
IJP PAN	Instytut Języka Polskiego Polskiej Akademii Nauk (Institute of Polish Language, Polish Academy of Sciences)
IS PAN	Instytut Slawistyki Polskiej Akademii Nauk (Institute of Slavic Studies, Polish Academy of Sciences)
LR	Language Resource/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
NASK	Naukowa i Akademicka Sieć Komputerowa (”Research and Academic Computer Network”, National Research Institute)
NLG	Natural Language Generation
NLP	Natural Language Processing
OPI	Ośrodek Przetwarzania Informacji (National Information Processing Institute, National Research Institute)
PAS	Polish Academy of Sciences
PCSS	Poznańskie Centrum Superkomputerowo-Sieciowe (Poznan Supercomputing and Networking Center)
PG	Politechnika Gdańska (Gdańsk University of Technology)
PJATK	Polsko-Japońska Akademia Technik Komputerowych (Polish-Japanese Academy of Information Technology)

PP	Politechnika Poznańska (Poznan University of Technology)
PŚ	Politechnika Śląska (Silesian University of Technology)
PW	Politechnika Warszawska (Warsaw University of Technology)
PWr	Politechnika Wrocławska (Wrocław University of Technology)
R&D	Research and Development
SR	Speaker Recognition
UAM	Uniwersytet im. Adama Mickiewicza (Adam Mickiewicz University)
UJ	Uniwersytet Jagielloński (Jagiellonian University)
UŁ	Uniwersytet Łódzki (University of Łódź)
UW	Uniwersytet Warszawski (University of Warsaw)
UWr	Uniwersytet Wrocławski (University of Wrocław)

Abstract

This study investigates the current (as of the first quarter of 2022) state-of-the-art language technology for Polish. Compared to two previous similar reports, created 10 and 5 years ago, the level of the support for Polish was much improved as a result of three independent trends.

The first one is Poland-specific and concerns some prominent events which have significantly changed the scope of operation of the Polish NLP community. They are: the construction of The National Corpus of Polish, the development of the CLARIN-PL infrastructure and an increase in funding of both scientific and R&D projects.

Two other trends are global. First, in recent years the development of speech and language resources and tools, previously carried out mostly by publicly funded institutions, was taken up also by private companies, often hiring scientists to their R&D departments. In the Polish context it created a valuable synergy, well observed in language technology evaluation campaigns such as PolEval.

Secondly, the deep learning revolution which is transforming our lives and economies, was also much reflected in the state-of-the-art for Polish. The release of freely available transformer models for Polish, based on Polish reference corpora, helped Polish benefit well in every natural language processing task.

This report highlights the advances made in language technology over the past 10 years and provides an insight to the current level of technology support available to Polish.

Streszczenie

Niniejsze opracowanie zarysowuje aktualny (zgodnie ze stanem z I kwartału 2022 r.) stan technologii językowej dla języka polskiego. W porównaniu z dwoma poprzednimi badaniami tego rodzaju, przeprowadzonymi 10 i 5 lat temu, poziom wsparcia dla języka polskiego uległ sporej poprawie, na co miały wpływ trzy czynniki.

Pierwszy z nich dotyczy kilku ważnych etapów rozwoju polskiej społeczności NLP, wyznaczonych przez powstanie Narodowego Korpusu Języka Polskiego, rozwój infrastruktury CLARIN-PL oraz wzrost finansowania projektów, tak naukowych, jak i badawczo-rozwojowych.

Dwa kolejne trendy mają charakter globalny. Po pierwsze, w ostatnich latach rozwojem zasobów i narzędzi mowy i języka, wcześniej realizowanym głównie przez instytucje finansowane ze środków publicznych, zajęły się także firmy prywatne, często zatrudniające naukowców do prac R&D. W polskim kontekście stworzyło to cenne synergie, dobrze widoczne np. w konkursie PolEval.

Po drugie, rewolucja związana z upowszechnieniem sztucznych sieci neuronowych, która zmienia nasze życie i gospodarkę, również znalazła swoje odzwierciedlenie w poprawie jakości technologii językowej dla polszczyzny. Udostępnienie modeli neuronowych dla języka polskiego na bazie danych z polskich korpusów referencyjnych wywarło wpływ na każde zadanie przetwarzania języka naturalnego.

Niniejszy raport zwraca uwagę na postęp, jaki dokonał się w technologii językowej w ciągu ostatnich 10 lat oraz przedstawia aktualny poziom wsparcia technologicznego dostępnego dla języka polskiego.

1 Introduction

The META-NET White Paper on Polish (Miłkowski, 2012) presented the current situation in the field of language processing in Poland, evaluating such factors as availability, quality and flexibility of technologies and resources in several areas based on an expert assessment. The report perfectly illustrated the state-of-the-art 10 years ago: indications of good support for speech synthesis were a result of the triumph of the Polish Ivona system in many international speech synthesis competitions, before its acquisition by Amazon, good quality of monolingual text corpora reflected the end of the multi-annual project of the National Corpus of Polish (Przepiórkowski et al., 2012) and the construction of the Polish Corpus of Wrocław University of Technology (Broda et al., 2012). At the same time the technologies for semantic analysis or text generation were categorised as having "weak or no support". The report contained mostly pessimistic conclusions: the lack of standardisation of resources, the need to intensify work on deep text analysis, ontological resources and electronic valency dictionaries.

Five years later a survey using categories from the report was distributed among representatives of the Polish LT community, resulting in the updated figures (Ogrodniczuk, 2017). The conclusions were optimistic e. g. for wordnets, corresponding to current projects (Maziarz et al., 2016) but much worse for the more advanced tools. For instance, deep semantic, pragmatic or discourse analytic tools were considered inaccessible, the quality of text summarisation was rated the weakest, together with word sense disambiguation, sentiment/opinion analysis, text generation and machine translation.

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

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and it seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – 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.

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

2 The Polish Language in the Digital Age

2.1 General Facts

Polish (from common Slavic core *pol* 'field') is the official language of the Republic of Poland and since 2004, the sixth largest official language of the European Union. It is spoken by 10% of EU citizens: about 40 million native speakers and 10 million second language speakers worldwide. In Poland, it is the common spoken and written language and the native language of the vast majority of the population.

[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.

Polish is a Slavic language of the Lechitic group, written in the Latin script. It is the most spoken West Slavic language in the world. It is quite homogenous, while the differences between its dialects (e. g. from Podhale region or the dialect of Poznań) are fairly small.

Particularities of the Polish language

Polish stands out with a large number of consonants (31 to 35) and a small number of vowels (6 to 8) depending on phonological status of palatal and nasal consonants and nasal vowels.

The language exhibits some specific characteristics, which contribute to the richness of the language (Pisarek, 2007) but present a challenge for computational processing. Word order is relatively free in Polish sentences, and it is used to stress the importance of information rather than simply follow from the rules of grammar. Polish is also relatively morphologically rich, which means that for roughly 180 thousand base forms of words, almost 4 million inflected word forms exist. The inflection paradigms are complex, and even their exact number is a matter of a dispute (single exceptions might be thought to create a new paradigm). Even native speakers have problems with properly inflecting many words, and most speakers of Polish as a second language, never completely master the complexities of the inflectional system.

Polish syntax is similar to its neighboring Slavic languages with a tendency to analyse constructions seen in gender marking, forms of address and the use of infinitive and impersonal constructions.

Language changes

The English language is one of the biggest sources of neologisms and calques, in particular in science and technology, and it exerts a considerable influence on contemporary Polish. The number of words loaned from English into Polish is however much lower than in Dutch or German because of the problem with inflecting some words and differences in pronunciation systems. In the early 1990s, just after the major political changes, companies used brands that sounded "English like". Even a grocery shop could bear an English signboard "*Your shop*". Today, such a name would be considered ridiculous by a much larger group of speakers. But calques from English, such as "*dokładnie*" (*exactly*) or "*wydawać się być*" (*seem to be*), are numerous and popular.

Another influence of English is the appearance of more direct forms of address – using the Polish pronoun "*ty*" ("*you*" singular) is quite popular these days. Arguably, this influence stems from incorrect, non-professional translations from English, yet it is a stable phenomenon. Similarly, Polish speakers are now more likely to follow English punctuation patterns, especially a comma after introductory phrase, which is, according to traditional Polish punctuation rules, incorrect. Even some typographical characters (such as "&"), never used in Polish before, are borrowed from English. One of the current developments in Polish is that feminine forms for professions are nowadays more frequently used, though they still remain somewhat outside of the official register.

Some of the traditional inflection patterns seem to undergo a process of simplification (for example, speakers are more likely to say "*mieliłem*" than "*mełłem*" (*I ground*), which would be the standard form), and some of the forms become almost extinct in everyday speech. This is especially true of the vocative case in colloquial Polish. Still, the inflection patterns are highly complex and no general trend towards simplifying them is discernible.

The vitality of Polish can best be seen in the annual poll *Youth Word of the Year* organized by Polish Scientific Publishers PWN in collaboration with the University of Warsaw. It is aimed at selecting the most popular words, expressions and phrases among young people in a given year. In 2021 the winning "*śpiulkolot*" (*sleeper*) was a joke word that meant a pleasant

place to sleep. The list of 20 words "nominated" by the jury reflected the tendency to give new meanings to words that were still present in the general language such as "*mrozi*" *it freezes*, now meaning *cringy*, high productivity of words with suffixes such as "–ara" or "-uwa" or ambiguous terms used to convey emotions such as *essa* (*great, more than OK*) or *sheesh* (*woe, yuck*).

Language protection and education

The protection of the Polish language is regulated by the Polish Language Act ("*Ustawa o języku polskim*") and is defined as e. g. taking care of the correct use of Polish, counteracting vulgarisation of the language, disseminating knowledge about the language and its role in culture, promoting respect for regionalisms and dialects and to prevent their disappearance, promotion of the Polish language in the world and supporting the teaching of the Polish language in Poland and abroad.

The Act established the Council for the Polish Language ("*Rada Języka Polskiego*") as an opinion-giving and advisory institution in matters of language use. At least every two years the Council presents to the parliament a report on the state of protection of the Polish language.

The Polish language is one of the most important subjects taught at all levels of education at a rate of 5 hours per week in elementary school and 4 hours in secondary school. Polish is also a compulsory subject for the school-leaving exam ("*matura*").

Children of Polish citizens residing abroad may study in Polish, or in the Polish language in schools and school consultation points, attached to diplomatic representations, consular offices and military representations of the Republic of Poland,in community Polish language schools or Polish sections in foreign schools. If they do not have such an opportunity in their country of residence, they may also learn Polish online in over 60 schools of the Centre for the Development of Polish Education Abroad ("*Ośrodek Rozwoju Polskiej Edukacji za Granicą*").

2.2 Polish in the Digital Sphere

In 2020, the number of .pl domains reached almost 2.5 million, decreasing from the previous year (the largest number was achieved in 2016 with over 2.7 million '.pl' domains).[2]

In 2021 the number of internet users in Poland amounted to 28.8 million, i. e., 87% of the population. The share of internet users who accessed the internet daily in 2020 was 72%.[3]

In February 2022 Polish Wikipedia ranked 11th in terms of the number of articles (currently over 1.5 million) and 10th in terms of the number of active users and edits.[4]

In 2021, the number of users of social networking sites in Poland amounted to three-quarters of internet users and more than half of adults now having accounts. In 2020, 92%t of Poles used YouTube in the past month. Facebook was used by 89% of Internet users, and Messenger by 72%. As of January 2022, Messenger was the most popular social media app among Polish users regarding the number of downloads.[5] The use of Polish in these media is predominant, with the tendency to use colloquial language, carelessness in relation to the use of punctuation marks etc. Most hashtags include Polish characters, but usually without case variation.

Most major general-use language applications have Polish interfaces but only some, with a notable exception of Google Assistant or DeepL, offer full support for Polish. Unfortunately Alexa and Siri still cannot communicate with Polish users in their native language.

[2] https://www.statista.com/statistics/1009848/poland-number-of-active-pl-domains/
[3] https://www.statista.com/topics/5573/internet-usage-in-poland/
[4] https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_edits_per_article
[5] https://www.statista.com/topics/5296/social-media-usage-in-poland/

3 What is Language Technology?

Natural language[6] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

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

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

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

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

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

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

LT is already fused in our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name a 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.[7]

4 Language Technology for Polish

4.1 Language Data and Tools

Two major factors have contributed to the development of language technology for Polish recent years. The first one is substantial funding available to researchers through two research infrastructures: CLARIN-PL and DARIAH-PL, helping develop the key language technology for Polish in a systematic way. The second factor is the deep learning breakthrough

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

which drives the development of the field today. The synergies created between language resources and tools result in modelling language in a generalised and unprecedented way. Together with the development of multi-task NLP benchmarks such as KLEJ[8] (Rybak et al., 2020) or PolEval evaluation campaign for NLP tools,[9] (Ogrodniczuk and Kobyliński, 2021) the Polish ecosystem of data, tools and evaluation procedures are on the cusp of significant improvements.

The ELE report places Polish in the similar position as most other EU official languages, with fragmentary support of most complex language processing tasks. However, BBasic text processing, e. g. part-of-speech tagging or syntactic parsing, seems to be well developed with many resources available[10] and accuracy scores comparable to their foreign counterparts.[11]

Several written corpora of contemporary Polish have been created, starting with the National Corpus of Polish, but due to copyright issues they cannot by freely available for download. This obstacle is often overcome with the use of freely distributable corpora such as KPWr (Polish Corpus of Wrocław University of Technology),[12] CCNet corpora,[13] Open Subtitles,[14] Wolne Lektury[15] or Wikimedia dumps.[16] Specialised corpora with CC-BY-accessible text also exist, with the largest one being the Polish Parliamentary Corpus.[17] (Ogrodniczuk, 2018) Written corpora of historical Polish are also being actively developed.[18]

The accessibility of multimodal spoken corpora and speech databases has also increased significantly in the last few years (Pęzik, 2018). One of the most recent developments in this area is the release of a large annotated corpus of phone-based customer support dialogs (DiaBiz).[19] The new resource is aimed at stimulating the development of dialog systems, including voicebots and conversational analytics solutions as a commercially viable domain of NLP, especially by small companies and academic research groups which have so far had to rely on very limited datasets.

A recent evaluation of major ASR services offered commercially for Polish conducted on the DiaBiz corpus (Pęzik and Adamczyk, 2022) shows that ASR Engines actively developed by Polish companies over the last decade are now on a par with solutions offered for this language by global NLP service providers. The best performing Poland-based providers evaluated in the report are known to have utilised some of the reference language resources described above including the National Corpus of Polish and multimodal speech corpora.

Goal-oriented chatbots and voicebots, usually DialogFlow or RASA-based, are getting more and more popular, with 40 active companies in the field and over 130 known deployments for big customers such as mBank, InPost, PKO, IKEA or DHL Poland.[20]

Many pre-trained word embeddings, language models and machine translation models for Polish have been made available recently,[21] including the two state-of-the-art transformer models: HerBERT (Mroczkowski et al., 2021) and plT5 (2021), created with the use of the National Corpus of Polish.

[8] https://klejbenchmark.com
[9] http://poleval.pl
[10] For example, the list of over 200 resources and tools for Polish at http://clip.ipipan.waw.pl/LRT.
[11] http://clip.ipipan.waw.pl/benchmarks
[12] http://nlp.pwr.wroc.pl/en/tools-and-resources/resources/kpwr
[13] https://github.com/facebookresearch/cc_net
[14] https://opus.nlpl.eu/OpenSubtitles-v2018.php
[15] https://wolnelektury.pl
[16] https://dumps.wikimedia.org
[17] http://clip.ipipan.waw.pl/PPC
[18] http://spxvi.edu.pl/korpus/, https://korba.edu.pl, http://korpus19.nlp.ipipan.waw.pl or http://chronopress.clarin-pl.eu.
[19] http://pelcra.pl/new/diabiz
[20] https://clutch.co/pl/developers/artificial-intelligence/chatbots and https://robonomika.pl/katalog-dostawcow-chatbotow-voicebotow-taskbotow-vi-edycja-luty-2022 (in Polish).
[21] https://github.com/sdadas/polish-nlp-resources

Still, such labour-intensive resources as multimodal corpora or corpora with discourse structure and discourse semantic annotations are practically unavailable.

4.2 Projects, Initiatives, Stakeholders

Language Technology in the Polish AI strategy

The main document describing the national Polish AI strategy is the *Policy for the Development of Artificial Intelligence in Poland from 2020* (Council of Ministers, 2020) published in September 2020. The document defines AI-related targets for Poland in the three perspectives: short-term (until 2023), medium-term (until 2027) and long-term (after 2027). They are additionally divided into 6 areas: AI and society, AI and innovative companies, AI and science, AI and education, AI and international cooperation and AI and the public sector. The development of natural language technology is mentioned as a short-term goal, supported with national grants for projects related to Polish language processing based on world-leading algorithms. Notably, the document mentions the importance of language data: the need for elimination of legal barriers to the exploration of language text corpora under copyright protection and awarding projects that make architecture, trained models and training data sets available for common use.

National research infrastructures supporting the Language Technology

The longest running research infrastructure in Poland related to language technology is CLARIN-PL[22] (Common Language Resources and Technology Infrastructure), developed in 2013 by a consortium consisting of six Polish research units with both technical and humanities profiles: Wrocław University of Technology (consortium leader), Institute of Computer Science of the Polish Academy of Sciences, the Institute of Slavic Studies of the Polish Academy of Sciences, Polish-Japanese Academy of Information Technology, University of Łódź and University of Wrocław. Over the years of scientific cooperation, the consortium has managed to create research teams, consisting of programmers, natural language processing specialists and linguists, specialised in creating technologies for the Polish language. The primary beneficiaries of the infrastructure are researchers, representatives of the humanities and social sciences (which stems from CLARIN-PL's commitments to CLARIN ERIC), but also representatives of other fields of science, public institutions and also commercial companies.

Another prominent Polish research infrastructure partially related to language technology is DARIAH-PL[23] (Digital Research Infrastructure for the Arts and Humanities), since 2014 the largest humanities consortium in Poland. The infrastructure enables acquisition, storage and integration of research data, diverse in form, content and provenance, as well as processing, visualisation and sharing of digital resources.

In 2020 both infrastructures were funded by the Polish Ministry of Education and Science and the European Regional Development Fund from the Programme of Development of modern research infrastructure of the science sector.

Recent changes in the Polish Language Technology landscape

There have been many developments, which have significantly changed the scope of operation of the Polish NLP community, the five most prominent being:

[22] https://clarin-pl.eu/index.php/en/home/
[23] http://dariah.pl/en/

1. The construction of the National Corpus of Polish[24] (NKJP; Przepiórkowski et al., 2012), a reference corpus containing over fifteen hundred millions of words sampled from very diverse sources and containing classic literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived and internet texts, balanced with respect to gender, age and regional distribution of samples. The availability of the corpus, and particularly its manually annotated 1-million word subcorpus, available on CC-BY licence, boosted both research in the humanities and development of many NLP tools, first of their kind for Polish. To name a few: morphological analyser Morfeusz[25] (Kieraś and Woliński, 2017), disambiguating tagger Concraft[26] (Waszczuk et al., 2018) or named entity recogniser Nerf[27] (Savary et al., 2010), some of historical value only, the other still in active use. Since the completion of the NKJP in 2011, other reference corpora have been used to represent recent developments in Polish. The most significant examples of such initiatives are the MoncoPL monitoring corpus (Pęzik, 2020) and The Corpus of the 2010s[28] currently under development by DARIAH-PL.

2. The development of the CLARIN-PL project,[29] a Polish part of the pan-European Common Language Resources & Technology Infrastructure aimed at researchers in the humanities and social sciences. The co-operation of many research institutions within the project led to the development of many NLP resources and tools such as Słowosieć, the Polish WordNet, a relational lexico-semantic dictionary of Polish[30] (Dziob et al., 2019), Korpusomat, a corpus creation tool for non-technical users[31] (Kieraś and Kobyliński, 2021), COMBO, a neural tagger, lemmatiser and dependency parser[32] (Klimaszewski and Wróblewska, 2021) or SpokesPL – a search engine for Polish conversational data[33] (Pęzik, 2015).

3. External funding in the form of grants, both European (Horizon 2020, Connecting Europe Facility) or national, distributed by the National Science Centre and National Centre for Research and Development, which allowed many research institutions and companies develop internal NLP solutions (often collaborating with each other). It should be noted that private companies have increasingly benefited from both public funding for industrial research (which often exceeds the typical budgets of basic research projects by an order of magnitude) and commercial demand for speech recognition or dialog systems. As a result, their NLP products are characterised by state-of-the art performance in such specialised areas.

4. The PolEval evaluation campaign for natural language processing tools for Polish[34] started in 2017 as a practical exercise intended to advance the state-of-the-art with a series of NLP tasks in which submitted tools compete against one another using available data and are pre-established evaluation procedures. The contest integrated the NLP community and revealed its structure (see Figure 1) and resulted in the development, enhancement and public release of reference datasets for NLP tasks such as sentiment analysis, speech recognition or machine translation.

[24] http://nkjp.pl
[25] http://morfeusz.sgjp.pl
[26] http://zil.ipipan.waw.pl/Concraft
[27] http://zil.ipipan.waw.pl/Nerf
[28] http://korpus-dekady.ipipan.waw.pl
[29] https://clarin-pl.eu
[30] http://plwordnet.pwr.wroc.pl/wordnet/
[31] https://korpusomat.pl
[32] https://github.com/360er0/COMBO
[33] http://spokes.clarin-pl.eu
[34] http://poleval.pl

5. The latest transformer models (HerBERT[35] and plT5[36]) are trained by the Machine Learning Research Team at Allegro and Linguistic Engineering Group at the Institute of Computer Science and the Polish Academy of Sciences, based on several large corpora of Polish, including the National Corpus of Polish. Making these models freely available for the community made enormous progress, increasing the quality and performance of many NLP applications.

The four pillars of Polish NLP

In recent years, the development of NLP resources and tools has moved from publicly funded institutions to private companies, often hiring scientists to their R&D departments. It is much the same in Poland and we can distinguish four intertwined communities in this process (see Figure 1):

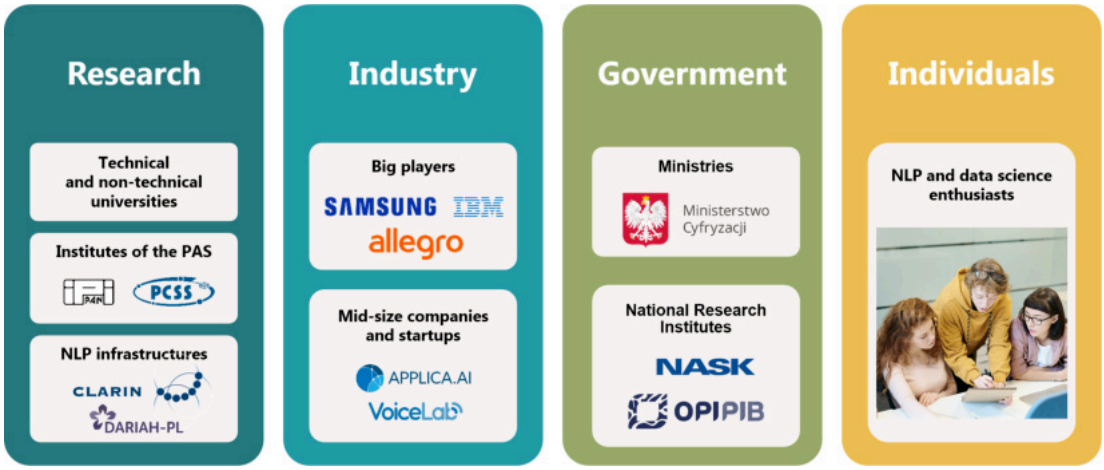

Figure 1: The four pillars of Polish NLP (with several example stakeholders)

1. Research labs and groups scattered around major Polish cities, located at the universities, both technical and non-technical and the institutes of the Polish Academy of Sciences (see Figure 2)
2. Government-based institutions and ministries, responsible for drafting strategic documents but also governing national research institutes involved in many NLP-intensive public projects
3. Companies, both the big international players as well as mid-size companies and startups
4. Independent researchers, without any formal affiliation, often forming informal research groups gathered around meetups.

[35] https://huggingface.co/allegro/herbert-large-cased
[36] https://huggingface.co/allegro/plt5-large

Figure 2: Main academic and public institutions involved in Polish NLP

5 Cross-Language Comparison

The LT field[37] as a whole has evidenced remarkable progress during the last few 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 never seen 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[38] 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)

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

[38] 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.

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

5.2 Levels of Technology Support

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

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in $\geq$3% and $<$10% of the ELG resources of the same type
3. **Moderate support**: the language is present in $\geq$10% and $<$30% of the ELG resources of the same type
4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[39]

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[40] 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

[39] 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.

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

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.[41]

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

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

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[42] 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 3 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

[41] 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.

[42] In addition to the languages listed in Table 1, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

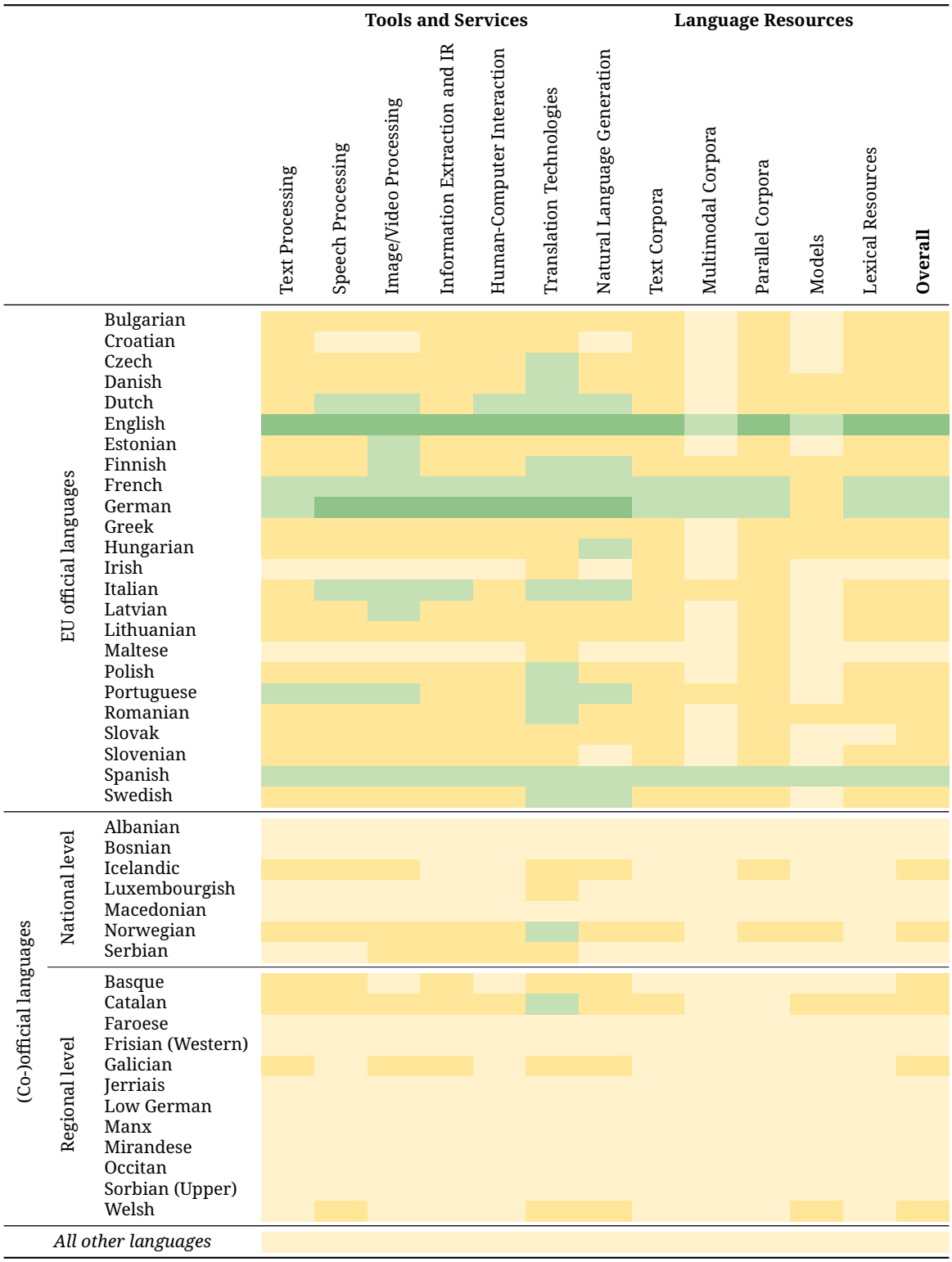

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

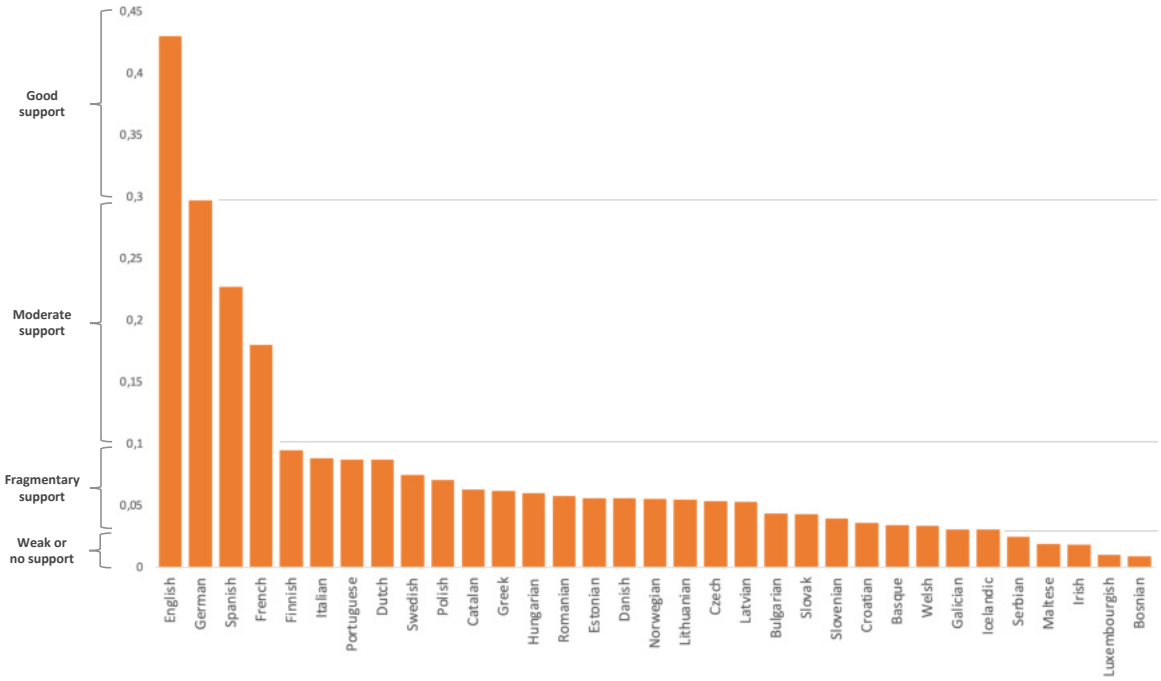

Figure 3: Overall state of technology support for selected European languages (2022)

on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

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

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

6 Summary and Conclusions

The current situation of Polish NLP seems to be in line with the development of the field in other European countries. Still, the high number of resources available in the existing

CLARIN-PL[43] or ELG repositories[44] does not imply their high adoption and rapid changes in the state-of-the-art require their constant update. Below we summarise the current situation and pinpoint suggested directions for the Polish LT research community, taking into account its local context.

1. Linking NLP to AI. This has already happened with the advent of deep neural network-powered solutions but its consequences are more far-reaching than the community can imagine. The NLP world became largely language-agnostic which moves the focus from low-level development of language-specific solutions to fine-tuning of existing models. But it also brings focus to advanced linguistic properties of individual languages and creating new synergies between linguists and engineers.

2. Increasing awareness of the value of data. The availability of deep neural network-powered frameworks moves the focus from tools to resources. This includes opening public data, already started with the Poland's Data Portal,[45] increasing access to trusted data and elimination of legal barriers to the exploration of Polish data under copyright protection.

3. Uptake of language technology. Even when the technology seems mature enough to be deployed commercially and is successfully adopted by startups, its take up by larger public institutions and companies is much slower. One step towards overcoming this barrier is the INFOSTRATEG programme[46] linking AI and NLP with direct practical applications and with GovTech-supported hackathons such as HackYeah[47] but it should be further supported by creating synergies between SMEs and research institutions.

4. Staying local in the globalising world. The globalisation of NLP creates enormous opportunities not just for Polish NLP labs which can compete with their colleagues abroad but also apply these developments to local contexts. In Poland it could mean intensification of NLP work for regional languages and dialects such as Kashubian or Silesian, still in their beginnings.

5. Support of the national research community with international co-operation. The Polish NLP research has already benefited from numerous pan-European initiatives such as ELRC, ELG and ELE, research infrastructures such as CLARIN and DARIAH, COST Actions and CEF projects. This trend must continue to strengthen the European research community.

6. Preparing for the future. The two most under-developed areas in Polish NLP seem to be discourse analysis and multimodality. Both require even more data, processing power and co-operation but only in this way can the process of understanding the human language, in all its complexity, be complete. All in all, interesting times are ahead!

References

plT5 Large — T5-based language model trained on Polish corpora. https://huggingface.co/allegro/plt5-large, 2021. Accessed: 2021-12-12.

[43] https://clarin-pl.eu/dspace/
[44] https://european-language-grid.eu
[45] https://dane.gov.pl
[46] https://www.gov.pl/web/ncbr/infostrateg
[47] https://hackyeah.pl

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.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. KPWr: Towards a Free corpus of Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3218–3222, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/965_Paper.pdf.

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

Council of Ministers. Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020 — The Policy for the development of AI in Poland from 2020, 2020.

Agnieszka Dziob, Maciej Piasecki, and Ewa Rudnicka. plWordNet 4.1 - a linguistically motivated, corpus-based bilingual resource. In *Proceedings of the 10th Global Wordnet Conference*, pages 353–362, Wroclaw, Poland, 2019. Global Wordnet Association. URL https://aclanthology.org/2019.gwc-1.45.

Witold Kieraś and Łukasz Kobyliński. Korpusomat – stan obecny i przyszłość projektu. *Język Polski*, CI(2):49–58, 2021. doi: https://doi.org/10.31286/JP.101.2.4. URL https://jezyk-polski.pl/index.php/jp/article/view/70.

Witold Kieraś and Marcin Woliński. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1):75–83, 2017.

Mateusz Klimaszewski and Alina Wróblewska. COMBO: State-of-the-art morphosyntactic analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 50–62, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.7. URL https://aclanthology.org/2021.emnlp-demo.7.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1213.

Marcin Miłkowski. *Język polski w erze cyfrowej – The Polish 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/polish. Georg Rehm and Hans Uszkoreit (series editors).

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.bsnlp-1.1.

Maciej Ogrodniczuk. Lingwistyka komputerowa dla języka polskiego: dziś i jutro (En. Natural Language Processing for Polish: today and tomorrow. *Język Polski*, XCVII(1):19–29, 2017. URL https://www.ceeol.com/search/article-detail?id=528569.

Maciej Ogrodniczuk. Polish Parliamentary Corpus. In Darja Fišer, Maria Eskevich, and Franciska de Jong, editors, *Proceedings of the LREC 2018 Workshop* ParlaCLARIN: Creating and Using Parliamentary Corpora, pages 15–19, Paris, France, 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-02-3. URL http://lrec-conf.org/workshops/lrec2018/W2/summaries/11_W2.html.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. *Proceedings of the PolEval 2021 Workshop*, Warsaw, Poland, 2021. Institute of Computer Science, Polish Academy of Sciences. ISBN 978-83-63159-31-3. URL http://poleval.pl/files/poleval2021.pdf.

Piotr Pęzik. Spokes — a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference*, pages 99–109, Linköpings universitet, 2015. Linköping University Electronic Press.

Piotr Pęzik. Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1678.

Walery Pisarek. *The Polish Language*. The Council for the Polish Language, 2007. ISBN 9788391626825. URL https://www.rjp.pan.pl/images/stories/pliki/broszury/jp_angielski.pdf.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. *Narodowy Korpus Języka Polskiego (En. National Corpus of Polish; in Polish)*. Wydawnictwo Naukowe PWN, Warsaw, 2012.

Piotr Pęzik. Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*, (7):133–150, 2020. doi: http://doi.org/10.31261/FL.2020.07.11. URL https://www.journals.us.edu.pl/index.php/FL/article/view/10335/7978.

Piotr Pęzik and Michał Adamczyk. Automatic Speech Recognition for Polish in 2022. An Intrinsic Evaluation of Selected ASR Engines on a Corpus of Customer Support Dialogs. Technical report, 2022. forthcoming.

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.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. KLEJ: Comprehensive Benchmark for Polish Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.111.

Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski. Towards the Annotation of Named Entities in the National Corpus of Polish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010. ELRA.

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

Jakub Waszczuk, Witold Kieraś, and Marcin Woliński. Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue: 21st International Conference (TSD 2018), Brno, Czech Republic, Proceedings*, number 11107 in Lecture Notes in Artificial Intelligence, pages 188–196. Springer-Verlag, 2018. ISBN 978-3-030-00794-2.