# EUROPEAN LANGUAGE EQUALITY

## D1.16

# Report on the German Language


| | |
|---|---|
| Authors | Stefanie Hegele, Barbara Heinisch, Antonia Popp, Katrin Marheinecke, Annette Rios, Dagmar Gromann, Martin Volk, Georg Rehm |



| | |
|---|---|
| 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 | D1.16 |
| Deliverable title | Report on the German Language |
| Type | Report |
| Number of pages | 25 |
| 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 | Stefanie Hegele, Barbara Heinisch, Antonia Popp, Katrin Marheinecke, Annette Rios, Dagmar Gromann, Martin Volk, Georg Rehm |
| Reviewers | Maria Giagkou, Sabine Kirchmeier |
| Editors | Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne |
| EC project officers | Susan Fraser, Miklos Druskoczi |
| Contact | European Language Equality (ELE)<br>ADAPT Centre, Dublin City University<br>Glasnevin, Dublin 9, Ireland<br>Prof. Dr. Andy Way – andy.way@adaptcentre.ie<br>European Language Equality (ELE)<br>DFKI GmbH<br>Alt-Moabit 91c, 10559 Berlin, Germany<br>Prof. Dr. Georg Rehm – georg.rehm@dfki.de<br>http://www.european-language-equality.eu |
| |  |

# Consortium

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

# Contents

# List of Figures



# List of Tables



# List of Acronyms

| | |
|---|---|
| ACDH-CH | Austrian Centre for Digital Humanities and Cultural Heritage |
| AI | Artificial Intelligence |
| ARD | Working group of public broadcasters of the Federal Republic of Germany (Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland) |
| BMBF | Federal Ministry of Education and Research |
| CL | Computational Linguistics |
| DACH | The region referring to Germany (D), Austria (A), and Switzerland (CH) |
| DFG | German Research Foundation |
| DLE | Digital Language Equality |
| ELE | European Language Equality *(this project)* |
| ELE Programme | European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)* |
| ELG | European Language Grid (EU project, 2019-2022) |
| EPFL | Swiss Federal Institutes of Technology in Lausanne |
| ETH | Swiss Federal Institutes of Technology in Zurich |
| FFG | Austrian Research Promotion Agency |
| FWF | Austrian Science Fund |
| GfDS | Society for the German Language |
| GPU | Graphic Processing Unit |
| GWK | Joint Science Conference (Gemeinsame Wissenschaftskonferenz) |
| HCI | Human Computer Interaction (see HMI) |
| HMI | Human Machine Interaction (see HCI) |
| HPC | High-Performance Computing |
| IDS | Institute for the German Language |
| LR | Language Resources/Resources |
| LT | Language Technology/Technologies |
| NFDI | National Research Data Infrastructure |
| NLG | Natural Language Generation |
| NLP | Natural Language Processing |
| OECD | Organisation for Economic Co-operation and Development |
| ÖAW | Austrian Academy of Sciences |
| PISA | Programme for International Student Assessment |
| SNF | Swiss National Science Foundation |
| SR | Speaker Recognition |

| | |
|---|---|
| UZH | University of Zurich |
| VDS | Verein Deutscher Sprache |
| WWTF | Vienna Science and Technology Fund |
| ZDF | Second German Television (Zweites Deutsches Fernsehen) |
| ZHAW | Zurich University of Applied Sciences |

## Abstract

With more than 150 million native and non-native speakers, German is the second most widely spoken language in the EU. It is a pluricentric language, with several interacting codified standard forms present in the region of Germany, Austria and Switzerland.

The last decade has seen strongly perceptible language change, trending towards the simplification of the grammatical system. With the internet being accessible to nearly everyone, the rise of social media has led to wider use of phenomena such as Anglicisms or emojis. In this context, the discussion of how to protect dialects from extinction, public debates about language policies fueled by right wing movements and whether standards for an inclusive gender-neutral language should be introduced, have gained a lot of attention. While change is omnipresent, the concern that German is seriously endangered because of Anglicisms, digitalisation or socio-political discussions can be dismissed, though. Several non-governmental institutions promote language protection and the study of German, not only in countries where it is an official language, but also abroad where it is one of the most popular studied second languages.

Due to its linguistic characteristics, the German language can be quite a challenge for natural language processing tasks. However, the list of language resources and language technologies for German is quite extensive. As of early 2022, there are approximately 2000 German data resources and tools listed in the ELG catalogue, with the actual number assumed to be significantly higher. Available resources include corpora, lexical/conceptual resources, language descriptions (i. e., computational grammars and language models) and tools.

While overall AI strategies vary in the German-speaking regions, the situation for language technology research and development in Germany is, all aspects considered, rather good. Germany has a thriving language technology industry. In addition to large corporations, there are many NLP start-ups located in Germany. More than 40 universities offer courses with topics related to language technology. Austria's current funding programmes do not focus directly on language technology. The number of courses that include language technology is comparatively small. Switzerland, on the other hand, has an active language technology industry including many start-ups and international technology companies. Swiss language technology has a stronger focus on multilingualism.

In order to withstand in the digital space, it is important for the German-speaking regions that incentives for research, digital education and also concrete opportunities for marketing and deploying LT applications are put in the forefront of future AI strategies.



## Zusammenfassung

Der vorliegende Bericht ist Teil einer Serie, die die gegenwärtige Lage der verschiedenen europäischen Sprachen beleuchtet. Dabei steht stets die Frage im Mittelpunkt: Welche Bedeutung hat die Digitalisierung von Informationen, Wissen und Kommunikation für die Sprache? Dieser Berichts geht sowohl auf allgemeine Entwicklungen und Besonderheiten der deutschen Sprache als auch auf ihre Unterstützung durch Sprachtechnologie ein.

Mit mehr als 150 Millionen Sprecher:innen ist Deutsch die am zweithäufigsten gesprochene Sprache in der EU und wird als pluriareale Sprache bezeichnet. In der DACH-Region (Deutschland, Österreich, Schweiz) werden nicht nur die drei kodifizierten Standardvarietäten des Deutschen gesprochen, sondern auch eine Fülle von Regionalsprachen und Dialekten. Zudem lernen über 15 Mio. Menschen weltweit Deutsch als Fremdsprache.

Der Sprachwandel, der sich vor allem in der rasant wachsenden Anzahl an Anglizismen, der abnehmenden Verbreitung von Dialekten und in soziopolitischen Debatten, wie der über die gender-neutrale Sprache manifestiert, hat einen erheblichen Einfluss auf den Sprachge-

brauch. In Bezug auf eine vielfach prophezeite Anglisierung des Deutschen oder einen weitgehenden Sprachverfall durch digitale Einflüsse kann jedoch Entwarnung gegeben werden. Unter anderem geben die Berichte zur Lage der deutschen Sprache zahlreiche Informationen und empirisch erhobene Daten über den Zustand der deutschen Sprache. Die Berichte werden von der Akademienunion und der Deutschen Akademie für Sprache und Dichtung herausgegeben und dienen der Meinungsbildung in der Öffentlichkeit sowie auch bei der politischen Entscheidungfindung, beispielsweise im Bildungsbereich.

Der Status einer Sprache hängt immer mehr von ihrer Präsenz im digitalen Informationsraum und den verfügbaren Softwareprodukten ab. Sprachtechnologie spielt hierbei eine entscheidende Rolle. Sprachtechnologie ist ein multidisziplinäres wissenschaftliches und technologisches Gebiet, das sich mit dem Studium und der Entwicklung von Systemen beschäftigt, die in der Lage sind, menschliche Sprachen zu produzieren und zu verstehen – sei es in geschriebener, gesprochener oder gebärdeter Form. Im vergangenen Jahrzehnt wurde die Sprachtechnologie insbesondere durch neuartige maschinelle Lernverfahren revolutioniert (Deep Learning).

Aufgrund zahlreicher linguistischer Besonderheiten bringt die deutsche Sprache eine Vielzahl von Herausforderungen für die maschinelle Verarbeitung natürlicher Sprache mit sich. Dennoch wird das Deutsche derzeit gut durch sprachtechnologische Produkte und Ressourcen unterstützt. Mit Stand Anfang 2022 sind in der Cloud-Plattform European Language Grid, die alle Sprachen Europas adressiert, etwa 2000 deutsche Datenressourcen und Services katalogisiert, wobei die tatsächliche Zahl deutlich höher liegen dürfte. Während die verfügbaren Sprachressourcen für das Deutsche insgesamt recht umfangreich sind, machen dialektspezifische Ressourcen zurzeit nur einen kleinen Prozentsatz aus. Für das Deutsche steht eine große Anzahl an Korpora zur Verfügung, zusammengestellt aus Zeitungsartikeln, Internetquellen oder sozialen Medien. Außerdem existieren mehr als 700 Tools und Anwendungen, sowohl für geschriebene als auch für gesprochene Sprache. Diese Ressourcen adressieren eine Vielzahl von Anwendungsbereichen, wie z. B. Sentimentanalyse, Themenklassifizierung, automatische Zusammenfassung, maschinelle Übersetzung und vieles mehr.

Um sich in der führenden Rolle im Bereich Sprachtechnologie zu behaupten, ist es für den deutschsprachigen Raum wichtig, Förderprogramme und Anreize für die Forschung zu schaffen. Die Situation für sprachtechnologische Forschung und Entwicklung in Deutschland ist verhältnismäßig gut, auch ohne dezidiertes Programm für die Entwicklung von Sprachtechnologien für die deutsche Sprache. Bis 2025 sollen drei Mrd. Euro unter anderem in den Aufbau neuer KI-Zentren, Förderprogramme, Professuren und internationale Kooperationen fließen. In diesem Zusammenhang existieren Leuchtturmprojekte wie z. B. OpenGPT-X, SPEAKER und QURATOR, die einen Fokus auf das Deutsche legen. Im letzten Jahrzehnt ist in Deutschland eine aufstrebende Sprachtechnologieindustrie entstanden. Neben großen Konzernen tragen auch vermehrt zahlreiche Startups im Bereich NLP und sprachbasierte KI zum Fortschritt bei. Mehr als 40 Universitäten bieten Studiengänge an, deren Inhalte unmittelbar oder mittelbar mit Sprachtechnologie zu tun haben. In Österreich ist die Anzahl an Studiengängen, die Sprachtechnologie beinhalten, zwar eher gering, doch existieren Förderprogramme, die sich, wenn auch indirekt, auf die Weiterentwicklung von Sprachtechnologie konzentrieren. Auch in der Schweiz spielt Sprachtechnologie in der Industrielandschaft eine immer größere Rolle, vertreten durch internationale Unternehmen und zahlreiche Startups.

# 1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support Europe's 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 additionally – and most importantly – to identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

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

This report has been developed by the European Language Equality (ELE) project[2] in the spirit of the META-NET White Paper Series *Europe's Languages in the Digital Age* (Rehm and Uszkoreit, 2012), especially with regard to the white paper on the German language (Burchardt et al., 2012). 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 German Language in the Digital Age

## 2.1 General Facts

With more than 150.000.000 native and non-native speakers (Eberhard et al., 2021), German is the second most widely spoken language in the European Union. German is an officially recognised language in seven European countries: Austria, Belgium, Germany, Italy, Liechtenstein, Luxembourg and Switzerland. German is considered a pluricentric, or rather pluriareal language (Scheuringer, 1996). Germany, Austria and Switzerland form the DACH region, which is not only home to the three (codified) standard varieties of the German language, but also boast a wealth of regiolects and dialects.

In Germany, the German language is the common spoken and written language as well as the native language of the vast majority of the population. Minority languages in the sense of the European Charter on Regional and Minority Languages include Danish and North Frisian in Schleswig-Holstein, Upper Sorbian in Saxony, Lower Sorbian in Brandenburg, Saterland Frisian in Lower Saxony, and the Romani language of the German Roma and Sinti throughout the country. Each group represents some tens to hundreds of thousands of speakers. In addition, there are immigrant languages, such as Turkish or Arabic (EFNIL European Federation of National Institutions for Languages, 2009).

The linguistic situation in the German speaking parts of Switzerland is a diglossia, where speakers use two varieties of German in everyday life. In formal contexts, people use Swiss Standard German ("Schweizer Hochdeutsch"), whereas in informal settings, Swiss German dialects are used ("Mundart"). The formal variety, Swiss Standard German, is relatively similar to Standard High German with some minor differences in grammar, orthography and vocabulary. By contrast, the informally spoken Swiss German dialects differ very substantially from Standard German. These spoken varieties are not uniform, but rather form a continuum within the (High and Highest) Alemannic dialect groups. Differences to Standard German include grammar (e. g., no preterite past), vocabulary (e. g., French loan words), and also phonology. Swiss German dialects are not easily intelligible to speakers of Standard German, if at all.[3]

[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://www.european-language-equality.eu

[3] To illustrate the differences, consider the following sentence in Standard German and Swiss German:
German: *Das war nicht Grossmutters Pferd, das wir dort drüben gesehen haben.*

The situation in Austria is similar to Switzerland with Standard German ("Hochdeutsch") being the codified form of German used in Austria. At the same time different regions boast a wealth of dialects. These can be attributed to different dialect regions, including several Bavarian ("Bairisch") regions and an Alemannic dialect region, close to the one in Switzerland. Speakers are used to code-switching between different German varieties depending on the context. The differences between the Standard German and Standard Austrian German are, among others, different pronunciation, different gender of nouns, the formation of compounds, the use of prepositions or tenses, syntax, and the most salient one, lexical differences (Wiesinger, 1996). A small set of Austriacisms (in the food domain) were mentioned in an annex to Austria's accession treaty (Protocol no. 10[4]) having special status in the EU legislation.

In general, the German language has many linguistic characteristics and particularities that pose a challenge to natural language processing tasks. Word order is relatively free. The compounding system allows for the combination of words and affixes in a simple way. As a consequence, there are many infinitely long German words. There is also a tendency to use fairly long, nested sentences. Separable verb prefixes can be positioned far away from their associated verb (Eroms et al., 2003).

As in many other languages, German uses a grammatical gender. However, nouns that are referring to the social gender are often biased towards the male form. Proponents of a gender-neutral language advocate that German needs a grammar that explicitly includes women and non-binary people, making all people feel equally addressed. A unified solution to the discussion sparked by Feminist Linguistics in the 1970s has not been found yet. One of the best-known solutions, called the gender star, puts an asterisk in front of the female word ending, such as Bürger*innen (citizens) (Kühne, 2017; Knoke, 2017).

Perceptible language change in German has been omnipresent for decades, leaving the language community to decide what becomes the norm. According to three reports on the state of the German language, published in the years 2013-2021 by the German Academy for Language and Poetry[5] and the Union of the German Academies of Sciences and Humanities[6], changes lean heavily towards the simplification of the grammatical system (Deutsche Akademie für Sprache und Dichtung and Union der deutschen Akademien der Wissenschaften, 2021). Also, the use of the conjunctive I and the genitive (Sick, 2004) are being increasingly displaced.[7] At the same time there has been a huge expansion in vocabulary. The German vocabulary has grown by more than 1.6 million words in the last 100 years (Eichinger et al., 2013). Over the last decades, the use of English in popular culture (television series, movies, music) has become prevalent, introducing many Anglicisms into the language that either replace existing German words or fill vocabulary gaps (Lemnitzer, 2007; Eichinger et al., 2013). Dialects have been more and more displaced. According to a recent study, only 57% of men and 50% of women in Germany still speak the "real old dialects" (Eroms, 2018). The change in Austrian dialects, which is sometimes perceived as dialect loss, is currently also a topic of debate among the speaker communities (Koppensteiner and Kim, 2020).

At present, there is no institutional body for official language protection in Germany. However, there are a number of non-governmental, publicly funded organisations that promote the study of German and encourage international cultural exchange. Among these institu-

Swiss: *Dasch nid am Grosi sis Ross gsi, womer det äne gseh händ.*
English: This was not Grandmother's horse that we saw over there.

4 http://data.europa.eu/eli/treaty/acc_1994/act_1/sign

5 https://www.deutscheakademie.de

6 https://www.akademienunion.de

7 This does not apply to the Swiss dialects, as those are deeply rooted in Swiss culture and have seen increased usage in recent years, not only in informal contexts but also in the media and by literary authors (Hollenstein and Aepli, 2015; Aepli and Clematide, 2018).

tions are the Goethe Institute[8], the Society for the German Language (GfDS)[9], the Institute for the German Language (IDS)[10] and Verein Deutsche Sprache (VDS)[11]. The Duden[12] lexicon (Werner, 2018) is the preeminent language resource of Standard High German and is updated regularly.

Public debates about language policy positions are becoming more frequent and also more heated. They attract a great deal of media attention in Germany. The New Right tries to use the topic of language in a targeted manner and to instrumentalise it in terms of national identity. A draft law provides for German to be codified as the national language in the German Constitution (Lobin, 2021). Compared to Germany, speakers in Austria are more lenient with Anglicisms or Teutonisms (Ransmayr, 2017).

Regarding language education, there has been some significant improvement in German students' reading literacy. According to the PISA study from 2018, performance is at a similar level to 2009, well above the initial results in 2000 and above the OECD average, especially among students from immigrant families (OECD, 2009; Bundesministerium für Bildung und Forschung, 2019). Unfortunately, the PISA study again confirms the strong correlation between socio-economic background and educational success (Avenarius et al., 2013; Bundesministerium für Bildung und Forschung, 2019; Beißwenger et al., 2021). Fears that the use of social media such as Twitter, Facebook, and Instagram would worsen young people's writing skills cannot be confirmed from a linguistic point of view. Rather, the emergence of new written forms should be noted (Storrer, 2014). The same goes for the increased use of emojis in digital text communication. Pictographs have changed how we use texts. Online conversations resemble oral conversation more and more (Beißwenger and Pappert, 2020). Moreover, digital text communication allows for the use of dialects also in written form. Although there is no codified form for Austrian dialects, for example, since they are usually only orally transmitted and used in oral communication, dialects are now also used more frequently in written communication.

German is currently the second most studied foreign language in the EU, but is also gaining in importance in Africa and Asia. The survey "German as a Foreign Language Worldwide" (Goethe Institut, 2020) shows that over 15.4 million people worldwide are learning German. The number of schools offering German language instruction has grown from 95.000 in 2015 to about 106.000 schools in 2020.

## 2.2 German in the Digital Sphere

German has a widespread online presence. It has the fourth largest Wikipedia (Wikimedia, 2021). Internet use continues to rise. According to a study by the public broadcasters ARD and ZDF, 94% of the German-speaking population in Germany over the age of 14 use the internet at least occasionally. This corresponds to 66.4 million of the total 70.6 million people aged 14 and over in Germany, an increase of 3.5 million people compared to 2019 (ARD/ZDF-Forschungskommission, 2021). Both, Germany and Austria have more than 85% of regular internet users and close to 70% of people with basic or above basic digital skills (Statistiken Österreich (Statistics Austria), 2011; Public Libraries, 2019; Eurostat, 2021).

[8] https://www.goethe.de
[9] https://gfds.de
[10] https://www.ids-mannheim.de
[11] https://vds-ev.de
[12] https://www.duden.de

# 3 What is Language Technology?

Natural language[13] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

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

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

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar(Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed 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 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.

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

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

- **Text Analysis** aims at identifying and labelling the linguistic information underlying any natural language text. 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.

[13] 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 digital devices using spoken language. Some of the main areas are Text to Speech Synthesis (TTS), i. e., the generation of speech given a piece of text, Automatic Speech Recognition (ASR), i. e., the conversion of a speech signal into text, and Speaker Recognition (SR).
- **Machine Translation** is the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** 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) 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.
- **Human-Computer Interaction** 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 into our day-to-day 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.[14]

# 4 Language Technology for German

This section provides a comprehensive and large-scale review study of the level of support the German language receives through Language Technology. The investigation in Sections 4.1 and 4.2 are based on a comprehensive metadata collection activity that aimed to collect, discover and appropriately document, ideally, all data sets, tools, services, components, repositories, companies, research groups etc. pertinent to LT for the German language. The data has been imported into the European Language Grid (ELG) platform and will facilitate

[14] In a recent report from 2021, the global LT market, already valued at USD 9.2 billion in 2019, 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).

computations of the Digital Language Equality metric (DLE) and comparative visualisations across languages, highlighting strong and weak points of digital support offered to each of the languages under investigation.

Even though the list of language resources and language technologies for German is quite extensive, it must be noted that the actual figure of existing resources is certainly higher. Despite an exhaustive search, there are resources and technologies that are currently neither documented not accessible. This is due to several reasons. First, researchers or developers might not be aware that they are creating language resources which might be useful and relevant for the field of Language Technology. Hence, they are only used internally by a small number of people. Second, developers may be aware of the value of their resources but do not distribute or document them publicly. This may be due to copyright reasons, confidentiality, (national) security reasons etc. The same applies to private sector companies.

Section 4.3 outlines the different national initiatives, projects, research structures and stakeholder groups for LT/AI in Germany, Austria and Switzerland.

As of January 2022, there are approx. 2000 German data resources and tools listed in the ELG catalogue. Currently, approx. 3% of all German data resources are tagged with the keyword Austrian German to indicate the geographic region. Less than 1% are tagged with Switzerland as a specific language geographical variety.

Resources can be divided into different categories including corpora, lexical/conceptual resources, language descriptions (i. e., grammars and language models) and tools.

## 4.1 Language Data

### Corpora

The current ELG catalogue lists a large number of German corpora of different sizes, ranging from a few hundred sentences up to million of sentences. The sources are most often newspaper texts or texts collected from the web and social media. Out of the more than 700 corpora almost half of them have additional linguistic information. Corpus annotations are a crucial enrichment for future research and development. Annotations can be of different nature and cover a large spectrum of syntactic, semantic, and discourse structure markup. The most common annotation types for the openly available German corpora are alignment, part of speech, lemma, sentiment and named entity. Corpora can either be multilingual, monolingual or bilingual. The distribution of the different types currently accessible is reasonably balanced (44% multilingual, 36% monolingual, 20% bilingual).

The vast majority, almost 75%, are text corpora. Compared to written language resources, the number of spoken corpora or multimedia corpora is relatively low. A quarter of all available corpora include one or more other media types such as audio, image and video. The collection includes a mix of very small corpora developed for a specific domain with carefully elaborated annotations as well as large corpora suitable for machine learning tasks. The most frequent corpora domains include health, news, politics and social media.

### Lexical conceptual resources and grammar/language models

Out of the approx. 400 lexical conceptual resources, more than half were classified as either terminological resources, lexica, dictionaries or word lists. Frequently assigned tags for the domains were the EU, law, politics and science. The vast majority of all lexical conceptual resources are of the media type text. More than half of the resources are multilingual, with English, French and Spanish being the other most commonly occurring languages.

Currently, there is only a small number of grammars/language models listed for German. Half of them are multilingual, the other half monolingual or bilingual. Subclasses that were

assigned for grammars/language models are: machine learning models, computational grammars, n-gram models, word embeddings and knowledge representation algorithm.

In addition, there are numerous free multilingual resources available online for German. The dictionary LEO[15] covers translation from German into eight other languages. Other widely used automatic translators are Deepl[16], which can translate German into 23 languages, and Google Translate[17], which covers the translation of 107 languages. EUROPEANA, Europe's Digital Library[18], launched in 2008, functions like a multimedia online portal with content from different sources. By the end of 2009, Germany, Austria and Switzerland had contributed around 16% to the more than 4.6 million objects (European Commission, 2009). Public Libraries 2030[19], PL2030, an international not-for-profit association, published in 2019 numbers on Europe's countries on digital skills.

## 4.2 Language Technologies and Tools

Currently, there are 700 tools that work either exclusively for German (35%) or multiple languages including German. The vast majority of tools take text as input for further processing. Even though speech technology has already been successfully integrated into many everyday applications, from spoken dialogue systems and voice-based interfaces to mobile phones and car navigation systems, audio is only supported by approx. one out of ten tools, image and video by even less. Research of the last decade and the deployment and integration of LT components to end-to-end processing pipelines has successfully led to the design of high-quality software with many tools supporting more than one function. The most frequent tasks supported by the current collection of German tools include text and data analytics, information extraction, named entity recognition, information retrieval and speech recognition. Tools developed by universities and research centres are usually available for all users free of charge. More than half of the tools currently listed are owned by companies that use commercial licenses for most of their products, which are typically available for a one-time or subscription fee. Some tools are available for free in their basic online demo versions, while upgrades require certain fees.

## 4.3 Projects, Initiatives, Stakeholders

The "AI Watch National strategies on AI: A European perspective in 2019" report (Van Roy et al., 2020) analyses the EU national AI strategies to identify areas for synergies and collaboration. While there is no LT-specific funding programme for Germany according to this report, the situation for LT research and development in Germany is rather good. Funding for LT-related topics is provided through research funds available for AI-related topics and also, on a more general level, through basic research support. In 2018, the government published its national AI strategy (Bundesregierung, 2018) which was updated in 2020 (Bundesregierung, 2020). Language analysis and understanding is (under the umbrella of HCI) one of five focus areas for innovation. The German government aims to invest approx. 3 billion Euro until 2025 to implement the strategy, including the creation of new AI centres, new funding programmes, new professorships, new international collaborations (e. g., with France) and a new national roadmap for AI standardisation.

For research and industry, these are additional opportunities on top of the established funding instruments (e. g., German Research Foundation, DFG[20], and Federal Ministry of Ed-

[15] https://dict.leo.org/englisch-deutsch/
[16] https://www.deepl.com/translator
[17] https://translate.google.com/?hl=de
[18] https://www.europeana.eu/de
[19] https://publiclibraries2030.eu
[20] https://www.dfg.de

ucation and Research, BMBF[21]). It remains to be seen if LT-related projects will rather focus on English (to be able to compete with the international scientific community that has been predominantly focusing upon English) or on German. The project SPEAKER (2020-2023) is an example of the latter category. It develops a conversational agent platform for the German language. Initiated by Germany and France, the GAIA-X[22] initiative works on the development of a federation of data infrastructure and service providers for Europe. A GAIA-X funding competition initiated the AI/LT project Open GPT-X, which started in January 2022, which develops large language models for German, English and a few other languages that will be shared with companies and research institutions in Europe. In 2021, the Joint Science Conference (Gemeinsame Wissenschaftskonferenz, GWK) decided to fund the NFDI for Data Science and Artificial Intelligence proposal under the umbrella of the Nationale Forschungsdateninfrastruktur initiative and programme (German National Research Data Infrastructure). NFDI4DataScience[23] will support all steps of the interdisciplinary research data life-cycle, including collecting/creating, processing, analysing, publishing, archiving and reusing resources in Data Science and Artificial Intelligence. Like its "companion" NFDI project Text+[24], NFDI4DataScience emphasises, among others, the handling and processing of language data.

Germany has a flourishing LT industry. At the point of writing, there are more than 100 developers and providers of LT software headquartered in Germany. They vary from small and medium enterprises focusing on one specific area to large and established companies such as SAP AG or Robert Bosch GmbH etc. In recent years, there has been a rise in NLP and language-centric AI startups (Startupill, 2021), with many success stories such as Explosion.AI, the developers of spaCy, a free, open-source library for advanced NLP. More than 40 universities in Germany offer LT-related subjects as part of their curricula. While some are more focused on linguistics, as part of the humanities faculties, others are situated at technical universities. Universities including, among others, Saarland University or the Institute for Natural Language Processing (IMS) at Stuttgart University have been performing world-class research in their respective fields.

Austria's AI strategy (Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie, 2021) addresses language technologies only indirectly. While aspects of language processing, speech recognition and voice control are mentioned as examples for AI applications, no concrete measures are targeting language technologies or resources. The same holds true for Austria's digital roadmap (Bundeskanzleramt und Bundesministerium für Wissenschaft, Forschung und Wirtschaft, 2016), which specifies that unstructured data, such as speech, are important for AI applications, but does not consider language data as part of big data. Although LRs and LTs are not explicitly mentioned in the strategy or roadmap, they certainly play a crucial role in the development of AI for the different areas of application listed in the Strategy, such as AI for the health sector, culture, media and education. Despite the fact that Austria has a dedicated language resource portal (*Sprachressourcenportal Österreichs*)[25], whose development is also described by Heinisch and Lušicky (2020), it does only contain a small number of language resources, and only one language technology, namely the EU Council Presidency Translator (Lušicky et al., 2019, unpublished manuscript). The major stakeholders behind the Austrian Language Resource Portal are the Centre for Translation Studies of the University of Vienna and the Language Institute of the Austrian Armed Forces. The EU Council Presidency in 2018 was a major driver for the development of the Portal, which also helped to increase the visibility of an informal working group in Austria's public administration, namely ARG GUT (Arbeitsgruppe Gouvernementaler Uebersetzungs- und Terminologiedienste).

[21] https://www.bmbf.de
[22] https://www.gaia-x.eu
[23] https://www.nfdi4datascience.de
[24] https://www.text-plus.org
[25] https://sprachressourcen.at

In Austria, funding programmes usually do not explicitly focus on LRs and LTs but rather on fundamental research, concrete applications or digitalisation. The Austrian Science Fund (FWF)[26] provides funding for fundamental research in any discipline. Therefore, the FWF does not fund applied or application-oriented research. The Vienna Science and Technology Fund (WWTF) with its Digital Humanism call[27] funds interdisciplinary projects by researchers from the social sciences, humanities and computer science to collaborate on the topic of human-centred technology that reflects social and humanistic values. This call is a reaction to the goals and intentions mentioned in the Vienna Manifesto on Digital Humanism (Werthner, 2019), which proclaims that human values must be reflected in technology. Research that fits the criteria mentioned above is also supported by the Austrian Research Promotion Agency (FFG), which is the national funding agency for industrial research and development. Digitalisation is also covered in Austria's funding landscape. The Austrian Academy of Sciences (ÖAW), especially its Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)[28], is a major stakeholder regarding the development and provision of LTs and LRs in Austria.

The number of Austrian universities that have a major language technology focused research strand or a dedicated degree programme is rather low. The Centre for Translation Studies at the University of Vienna is engaged in different initiatives, such as ELRC, ELE or NexusLinguarum (addressing linguistic data science) and led one ELG pilot project (Wachowiak et al., 2021).

In Switzerland, the main source of funding for LT comes from the Swiss National Science Foundation (SNF) and the Swiss Innovation Agency (InnoSuisse) – the former has a strong emphasis on academic research across all fields, while the latter provides funding for projects with industry partners and has generally a stronger focus on practical applications. Numerous universities across the country have dedicated LT groups and offer degree programs in Computational Linguistics or related fields such as Digital Humanities, including the Swiss Federal Institutes of Technology in Zurich (ETH) and Lausanne (EPFL), the University of Zurich (UZH), the Zurich University of Applied Sciences (ZHAW) and more. Switzerland has an active LT industry that includes local engineering hubs of international companies such as Google and Facebook, but also many small startups that cover a wide range of NLP applications. In general, language technology in Switzerland has a strong focus on multilinguality, both in terms of applications and resources. Furthermore, LT for the local Swiss German dialects has received more attention in the past years with research projects (e. g., "What's up, Switzerland?"[29]), resources (e. g., SwissDial dataset[30] and applications (e. g., slowsoft's TTS for Swiss German varieties[32]).

# 5 Cross-Language Comparison

The LT field[33] 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

[26] https://www.fwf.ac.at/en
[27] https://www.wwtf.at/digital_humanism
[28] https://www.oeaw.ac.at/acdh/acdh-ch-home
[29] https://whatsup-switzerland.ch/index.php/en/
[30] https://mtc.ethz.ch/publications/open-source/swiss-dial.html, Idiotikon[31]
[32] https://slowsoft.ch/eng/products.html
[33] This section has been provided by the editors.

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[34] 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.)

## 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

[34] 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.

3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[35]

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[36] 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 development of a 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.[37]

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

[35] 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.

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

[37] 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.

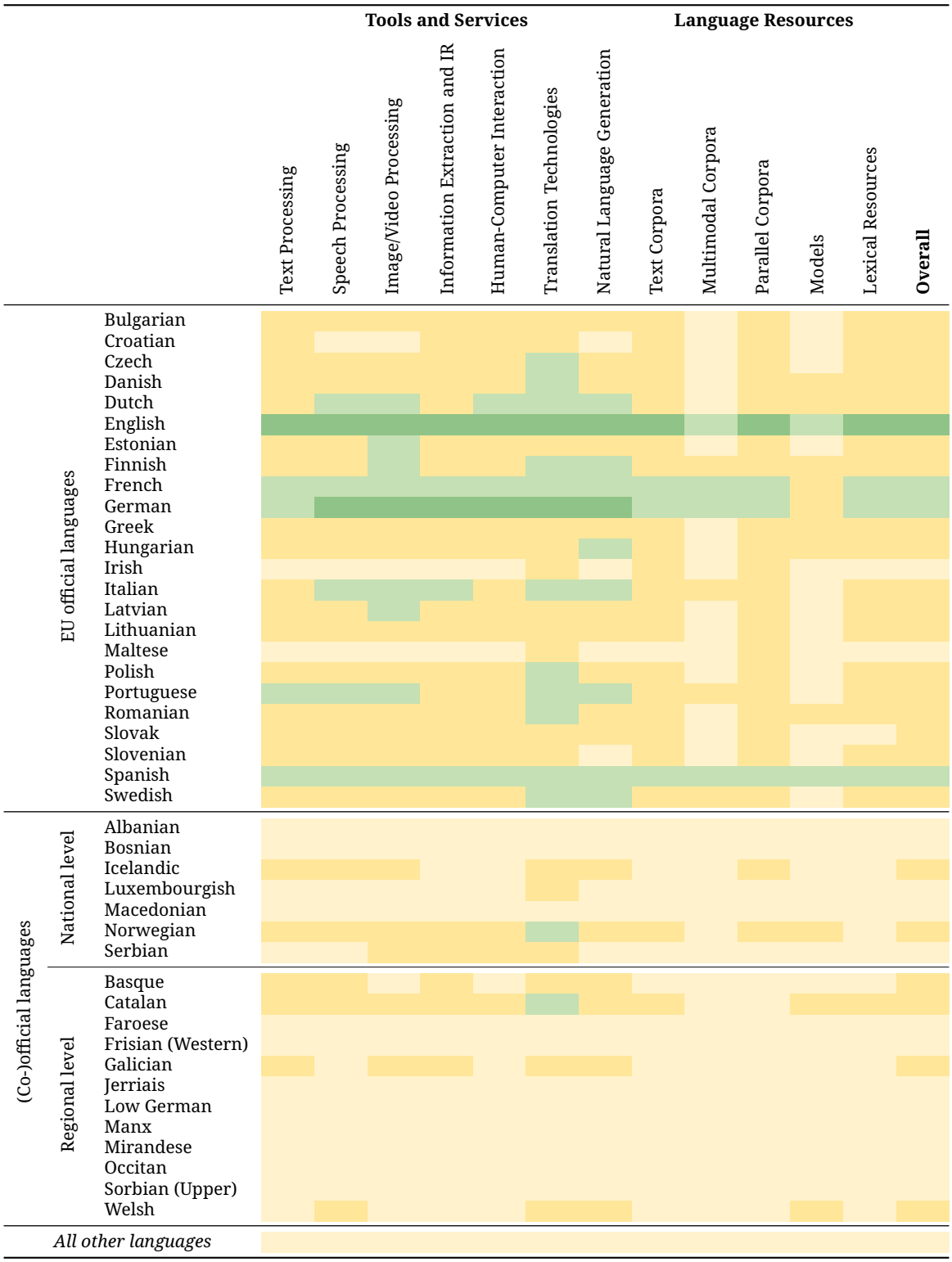


Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

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,[38] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

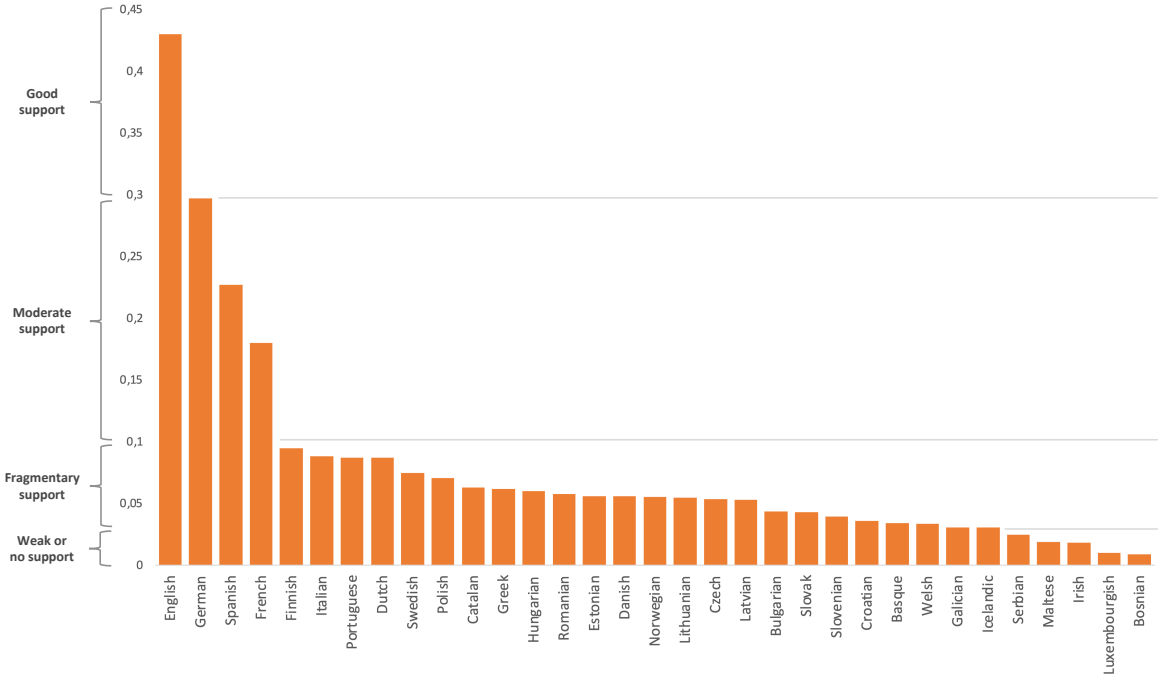


Figure 1: Overall state of technology support for selected European languages (2022)

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and im-

[38] 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.

balance, 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

German is the most widely used language in the EU after English. The availability of many different corpora and tools shows that German is currently well supported through language technologies, even though the many linguistic particularities make German a tough nut to crack for many LT tasks. A high number of large-scale resources and state-of-the-art technologies have been produced for Standard German. However, the scope of the resources and the range of tools are still limited when compared to English, and they are not yet good or ample enough to develop the kind of technologies required to support a truly multilingual knowledge society. Currently, existing technologies do not cover the many varieties of regional languages and dialects that exist in Germany, Austria and Switzerland.

Language varieties and also dialects deserve particular attention in the German language. Since Germany has roughly ten times the number of inhabitants of Austria or Switzerland, it also produces the highest number of language data and language resources. While the standard varieties are supported by language technologies, at least to some extent, non-standard varieties of the German language are often not taken into consideration (exceptions are language resources that 'translate' between the standard varieties of the German language).

The research community in Germany, Austria and Switzerland has been growing rapidly over the last decade. Numerous universities offer study programmes focused on LT, NLP, CL and closely related disciplines such as Digital Humanities or Applied Linguistics. Recent breakthroughs in AI have not only led to cutting edge technology developed by big companies, but have also also inspired the foundation of various startups and SMEs in the field. Current funding programmes, even though mostly targeted towards AI, have also helped to improve research in the field in general, and also have supported a number of research projects working on German in particular.

Recent studies have shown that Language Technology still has lots of untapped potential. For instance it can help provide students with impeded language skills, by making their learning experience more attractive. More generally, language technologies such as automatic speech recognition, speech synthesis, text analysis, language generation or machine translation can help reduce many disadvantages existing in our society. Assistive technologies already facilitate various tasks for persons with disabilities, but needs to be improved further (Ebling, 2018).

# References

Noëmi Aepli and Simon Clematide. Parsing approaches for swiss german. In Mark Cieliebak, Don Tuggener, and Fernando Benites, editors, *Proceedings of the 3rd Swiss Text Analytics Conference (SwissText)*, volume 2226 of *CEUR Workshop Proceedings*, pages 6–16, June 2018. URL https://doi.org/10.5167/uzh-159152.

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.

ARD/ZDF-Forschungskommission. ARD/ZDF-Onlinestudie 2020: Zahl der Internetnutzer wächst um 3,5 Millionen, 2021. URL https://www.ard-zdf-onlinestudie.de/ardzdf-onlinestudie/pressemitteilung/. Pressemitteilung.

Hermann Avenarius, Hartmut Ditton, Hans Döbert, Klaus Klemm, Eckhard Klieme, Matthias Rürup, Heinz-Elmar Tenorth, Horst Weishaupt, and Manfred Weiß. *Bildungsbericht für Deutschland: Erste Befunde*. Springer-Verlag, 2013.

Michael Beißwenger and Steffen Pappert. Sprachverfall durch Emojis? Eine pragmalinguistische Perspektive auf den Beitrag von Bildzeichen zur digitalen Kommunikationskultur. *Aptum.Zeitschrift für Sprachkritik und Sprachkultur*, 16:32–50, 2020.

Michael Beißwenger, Kristian Berg, Dirk Betzel, Ursula Bredel, Helmuth Feilke, Vivien Heller, Katrin Kleinschmidt-Schinke, Miriam Langlotz, Beate Lütke, Moti Mathiebe, Miriam Morek, and Jonas Romstadt. *Die Sprache in den Schulen - Eine Sprache im Werden: Dritter Bericht zur Lage der deutschen Sprache*. Erich Schmidt Verlag, Berlin, 2021.

Bundeskanzleramt und Bundesministerium für Wissenschaft, Forschung und Wirtschaft. Digital Roadmap Austria. https://www.digitalroadmap.gv.at/fileadmin/downloads/digital_road_map_broschuere.pdf, 2016.

Bundesministerium für Bildung und Forschung. PISA 2018: Deutschland stabil über OECD-Durchschnitt, 2019. URL https://www.bmbf.de/bmbf/shareddocs/pressemitteilungen/de/pisa-2018-deutschland-stabil-ueber-oecd-durchschnitt. Pressemitteilung 149/2019.

Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie. Strategie der Bundesregierung für Künstliche Intelligenz. Artificial Intelligence Mission Austria 2030 (AIM AT 2030). Technical report, BMK, 2021. URL https://www.bmdw.gv.at/Themen/Digitalisierung/Strategien/Kuenstliche-Intelligenz.html.

Die Bundesregierung. Strategie Künstliche Intelligenz der Bundesregierung. *Berlin, November*, 2018.

Die Bundesregierung. Strategie Künstliche Intelligenz der Bundesregierung – Fortschreibung 2020. *Berlin, November*, 2020.

Aljoscha Burchardt, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im digitalen Zeitalter – The German 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/german. Georg Rehm and Hans Uszkoreit (series editors).

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

Deutsche Akademie für Sprache und Dichtung and Union der deutschen Akademien der Wissenschaften. *Die Sprache in den Schulen - eine Sprache im Werden : dritter Bericht zur Lage der deutschen Sprache*. Berlin, 2021. ISBN 9783503205035.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the World, 2021. Online Version: http://www.ethnologue.com.

Sarah Ebling. Sprachtechnologie für Menschen mit Behinderungen. Sonderpädagogik in der digitalisierten Lernwelt. Beiträge der nationalen Tagung Netzwerk Forschung Sonderpädagogik, pages 29–46, Zürich, 2018.

EFNIL European Federation of National Institutions for Languages. About Germany (Deutschland). http://www.efnil.org/documents/language-legislation-version-2007/germany/germany, 2009.

Ludwig Eichinger, Peter Eisenberg, Wolfgang Klein, Angelika Storrer, et al. Reichtum und Armut der deutschen Sprache: Erster Bericht zur Lage der deutschen Sprache. Herausgegeben von der Deutschen Akademie für Sprache und Dichtung und der Union der deutschen Akademien der Wissenschaften, 2013.

Hans-Werner Eroms. Vielfalt und Einheit der deutschen Sprache (2017). Zweiter Bericht zur Lage der deutschen Sprache. Herausgegeben von der Deutschen Akademie für Sprache und Dichtung und der Union der deutschen Akademien der Wissenschaften, 2018.

Hans-Werner Eroms, Gerhard Stickel, and Gisela Zifonun. Schriften des Instituts für Deutsche Sprache. 2003.

European Commission. EUROPEANA – Europe's Digital Library: Frequently Asked Questions, 8 2009. URL https://ec.europa.eu/commission/presscorner/detail/en/MEMO_09_366. Memo/09/366.

Eurostat. Individuals' Level of Digital Skills. Technical report, 2021. URL https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_sk_dskl_i&lang=en.

Goethe Institut. Deutsch als Fremdsprache weltweit 2020, 2020. URL https://www.goethe.de/de/spr/eng/dlz.html.

Barbara Heinisch and Vesna Lušicky. The Austrian language resource portal for the use and provision of language resources in a language variety by public administration – a showcase for collaboration between public administration and a university. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 28–31, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-62-7. URL https://aclanthology.org/2020.lt4gov-1.5.

Nora Hollenstein and Noëmi Aepli. A Resource for Natural Language Processing of Swiss German Dialects. In Bernhard Fisseni, Bernhard Schröder, and Torsten Zesch, editors, *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 108–109, October 2015. URL https://doi.org/10.5167/uzh-174601.

Mareike Knoke. Wie »gender« darf die Sprache werden? *Spektrum*, 2017.

Wolfgang Koppensteiner and Agnes Kim. *Perspectives on Change: Language (Varieties) Contact and Language Ideologies on German in Austria*, pages 317–358. 2020. doi: 10.14220/9783737011440.317. URL https://www.vr-elibrary.de/doi/abs/10.14220/9783737011440.317.

Anja Kühne. Sprachentwicklung – Mein Deutsch, dein Deutsch. *Der Tagesspiegel*, 2017.

Lothar Lemnitzer. *Von Aldianer bis Zauselquote: neue deutsche Wörter, wo sie herkommen und wofür wir sie brauchen*. Gunter Narr Verlag, 2007.

Henning Lobin. *Sprachkampf – Wie die Neue Rechte die deutsche Sprache instrumentalisiert*. Dudenverlag, Berlin, 3 2021.

Vesna Lušicky, Barbara Heinisch, and Matīss Rikters. The EU Council Presidency Translator – a Neural Machine Translation System for the EU Council Presidency: Collecting Training Data, Training of the Engine and Dissemination in Austria. Technical report, Centre for Translation Studies, University of Vienna, Austria, 2019, unpublished manuscript.

OECD. Summary of Results from PISA 2009 (PISA 2009 Ergebnisse: Zusammenfassung), 2009. URL http://www.pisa.oecd.org/dataoecd/34/19/46619755.pdf.

Public Libraries. Libraries and Skills in Germany. Technical report, 2019. URL https://publiclibraries2030.eu/wp-content/uploads/2019/12/Germany2019.pdf.

Jutta Ransmayr. Insiders' and outsiders' views on German from Austria's perspective: Austrian Standard German and German Standard German–the odd couple. *Stereotypes and linguistic prejudices in Europe*, pages 187–206, 2017.

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.

Hermann Scheuringer. Das Deutsche als pluriareale Sprache: Ein Beitrag gegen staatlich begrenzte Horizonte in der Diskussion um die deutsche Sprache in Österreich. *Die Unterrichtspraxis/Teaching German*, pages 147–153, 1996.

Bastian Sick. *Der Dativ ist dem Genitiv sein Tod – Ein Wegweiser durch den Irrgarten der deutschen Sprache (The Dative is the Genitive its Death – How to navigate the Labyrinth that is the German Language).* Kiepenheuer und Witsch, 2004.

Startupill. 51 Best Natural Language Processing Startups in Germany of 2021, 2021. URL https://startupill.com/51-best-natural-language-processing-startups-in-germany-of-2021.

Statistiken Österreich (Statistics Austria). Ikt nutzung in haushalten (ict usage in households), 2011. URL http://www.statistik.at/web_en/statistics/information_society/ict_usage_in_households/041019.html.

Angelika Storrer. Sprachverfall durch internetbasierte Kommunikation? In Albrecht Plewnia and Andreas Witt, editors, *Sprachverfall?*, pages 171–196. De Gruyter, Berlin, 2014. DOI https://doi.org/10.1515/9783110343007.171.

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.

Vincent Van Roy et al. AI Watch-National strategies on Artificial Intelligence: A European perspective in 2019. Technical report, Joint Research Centre (Seville site), 2020.

Lennart Wachowiak, Christian Lang, Barbara Heinisch, and Dagmar Gromann. Towards Learning Terminological Concept Systems from Multilingual Natural Language Text. In Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch, editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 22:1–22:18, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-199-3. doi: 10.4230/OASIcs.LDK.2021.22. URL https://drops.dagstuhl.de/opus/volltexte/2021/14558.

Jürgen Werner. Duden. Die deutsche Rechtschreibung. In *Forum Classicum*, number 1, pages 62–63, 2018.

Edward A. Lee Werthner, Hannes. Vienna Manifesto on Digital Humanism, May 2019. URL https://www.informatik.tuwien.ac.at/dighum/wp-content/uploads/2019/07/Vienna_Manifesto_on_Digital_Humanism_EN.pdf.

Peter Wiesinger. Das österreichische Deutsch als eine Varietät der deutschen Sprache. *Die Unterrichtspraxis / Teaching German*, 29(2):154–164, 1996. ISSN 0042062X, 17561221. URL http://www.jstor.org/stable/3531825.

Wikimedia. List of Wikipedias, 2021. URL https://meta.wikimedia.org/wiki/List_of_Wikipedias. Website.