Language Report German

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

DOI: https://doi.org/10.1007/978-3-031-28819-7_18

Chapter 18
Language Report German

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

Abstract German is the second most widely spoken language in the EU. The last decade has seen strongly perceptible language change, trending towards the simplification of the grammatical system, a rapidly growing number of anglicisms, a decreasing prevalence of dialects, and an increase in socio-political debates on matters such as language policies and gender-neutral language. Many technologies and resources for German are available, which is also due to numerous well-established research institutions and a thriving Language Technology (LT) and Artificial Intelligence (AI) industry. In order to withstand in the digital sphere, it is important that incentives for research, digital education and also concrete opportunities for marketing and deploying LT applications are put at the forefront of future AI strategies.

1 The German Language

With more than 150 million native and non-native speakers (Eberhard et al. 2021), German is the second most widely spoken language in the European Union. 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 boasts a wealth of regiolects and dialects. 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,[1] published in the years 2013-2021 by the Union of the German Academies of Sciences and Humanities, changes lean heavily towards the simplification of the grammatical system.

Stefanie Hegele · Katrin Marheinecke · Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany,
stefanie.hegele@dfki.de, katrin.marheinecke@dfki.de, georg.rehm@dfki.de

Barbara Heinisch · Dagmar Gromann
University of Vienna, Austria, barbara.heinisch@univie.ac.at, dagmar.gromann@univie.ac.at

Antonia Popp · Annette Rios · Martin Volk
University of Zurich, Switzerland, popp@cl.uzh.ch, rios@cl.uzh.ch, volk@cl.uzh.ch

[1] https://www.akademienunion.de/publikationen/sammelbaende

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_18

There has also been a huge expansion in vocabulary. Over the last decades, many Anglicisms have been introduced into the language, that either replace existing German words or fill vocabulary gaps. Dialects have been more and more displaced.

German uses grammatical gender. However, nouns that refer to the social gender are often biased towards the male form. Proponents of a gender-inclusive language advocate that German needs a grammar that explicitly includes women and non-binary people, making all people feel equally addressed.

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 (Lobin 2021).

There are a number of non-governmental, publicly funded organisations that promote the study of German and encourage international cultural exchange, such as the Goethe Institute, the Society for the German Language, or the Institute for the German Language.[2]

Regarding language education, the PISA study has continued to confirm the strong correlation between socio-economic background and educational success.[3] Fears that the increased use of social media and emojis would worsen young people's writing skills cannot be confirmed. Instead, the emergence of new written forms should be noted (Beißwenger and Pappert 2020; Storrer 2014).

German is currently the second most studied foreign language in the EU, but is also gaining importance in Africa and Asia.

German has a widespread online presence and the fourth largest Wikipedia. Internet use continues to rise. According to the European Statistical Office (Eurostat), in both Germany and Austria, there are more than 85% of regular internet users and close to 70% of people with basic or above basic digital skills.

2 Technologies and Resources for German

German has many linguistic characteristics and particularities such as relatively free word order and fairly long nested sentences (Eroms et al. 2003) that pose challenges for Natural Language Processing tasks. Nevertheless, German is well supported by Language Technology (LT) applications and resources compared to most other European languages. A number of large-scale resources and state-of-the-art technologies have been produced for Standard German. However, dialect-specific resources currently account for only a small percentage.

There exist a large number of German corpora of different sizes, ranging from a few hundred sentences up to millions. The sources are most often newspaper texts or texts collected from the web and social media. Various terminological resources, lexica, dictionaries or word lists have also been developed for German. Annotations

[2] https://www.goethe.de, https://gfds.de, https://www.ids-mannheim.de

[3] https://www.bmbf.de/bmbf/shareddocs/pressemitteilungen/de/pisa-2018-deutschland-stabil-ueber-oecd-durchschnitt.html

cover a large spectrum of syntactic, semantic, and discourse structure markup. The most frequent corpus domains include health, news, politics and social media. Currently, there are only a few language models publicly available for German.

In addition, there are numerous free multilingual resources available online for German, e. g., the LEO dictionary. Other widely used MT systems are DeepL and Google Translate which cover the translation from German into dozens of languages. EUROPEANA functions like a multimedia portal and digital library with content from different sources.[4] By the end of 2015, Germany, Austria and Switzerland had contributed around 16% to the more than 24 million objects.

Hundreds of tools, both open source and commercial, that work either exclusively for German or multiple languages including German have been developed. The vast majority process text input. 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 a small fraction of tools, and image and video by even less.

Research over 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 typically available for all users free of charge.

The research community in Germany, Austria and Switzerland has been growing rapidly over the last decade. Numerous universities offer study programmes focused on Language Technology, NLP, Computational Linguistics and closely related disciplines. Recent breakthroughs in AI have not only led to cutting-edge technology developed by big companies, but have also inspired numerous 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. While overall AI strategies vary in the German-speaking regions, the situation for LT/NLP research and development in Germany is, all aspects considered, rather good. The German government aims to invest about 3 billion Euros 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.

3 Recommendations and Next Steps

The scope of resources and 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 re-

[4] https://www.europeana.eu/de

quired to support a truly multilingual knowledge society. High quality data sets and large language models represent a major step forward in AI. Our empirical results show that German is still partially lagging behind in this area (Hegele et al. 2022; Burchardt et al. 2012). There are also gaps in the areas of speech and text processing. In addition, existing technologies do not cover the many different varieties of regional languages and dialects that exist in Germany, Austria and Switzerland. Furthermore, many resources are not available due to copyright reasons, confidentiality, (national) security reasons etc.

While German is among the three best supported European languages (next to Spanish and French), the gap towards English is indeed significant. Without a substantial and timely intervention by the European Union, for many European languages this gap will continue to increase, endangering their digital existence.

References

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

Burchardt, Aljoscha, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk (2012). *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. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/german.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (2021). *Ethnologue: Languages of the World*. Dallas, Texas. http://www.ethnologue.com.

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

Hegele, Stefanie, Barbara Heinisch, Antonia Popp, Katrin Marheinecke, Annette Rios, Dagmar Gromann, Martin Volk, and Georg Rehm (2022). *Deliverable D1.16 Report on the German Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-german.pdf.

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

Storrer, Angelika (2014). "Sprachverfall durch internetbasierte Kommunikation?" In: *Sprachverfall?* Ed. by Albrecht Plewnia and Andreas Witt. Berlin: De Gruyter, pp. 171–196.