Chapter 13
Language Report English

Diana Maynard, Joanna Wright, Mark A. Greenwood, and Kalina Bontcheva

Abstract This chapter focuses on the status of the English language, primarily acting as a benchmark for the level of technological support that other European languages could receive (see Maynard et al. 2022; Ananiadou et al. 2012). While it is rather unlikely that any other European language will ever reach this level, due to the continuing development of support for English, and thus serves as a moving goalpost, nevertheless it provides a good criterion for relative assessment. While the inequalities in the amount of technological support available for English compared with other European languages may act as a deterrent for working on the latter, nevertheless it serves as a useful mechanism for applying cross-lingual transfer methods in order to build language models and generate labelled data for lower resource languages.

1 The English Language

English is a truly international language, due in no small part to the worldwide influence of the British Empire since the 17th century, and later to the influence of the United States. It has become the primary language of international discourse and is the lingua franca in many professional contexts, as well as in a number of regions with diverse native languages. English is the most spoken language in the world, with an estimated 1.36 billion total speakers.English is also the most widely taught foreign language in the world. There are almost three times as many people who speak English as a second language compared to native speakers, with a total of 360 million first language speakers and around one billion second language speakers.

English is an Indo-European language and shares a number of features of other Germanic languages. It uses the Latin alphabet with a left-to-right writing system, and has the ISO-639-1 code (*en*). It is classed as a pluricentric language, meaning that it has no single standard codified form but rather several interacting ones, typically set by or corresponding to different countries (e. g., US vs. British English).

Diana Maynard · Joanna Wright · Mark A. Greenwood · Kalina Bontcheva
University of Sheffield, United Kingdom, d.maynard@sheffield.ac.uk, j.wright@sheffield.ac.uk, m.greenwood@sheffield.ac.uk, k.bontcheva@sheffield.ac.uk

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_13

English is the most commonly used language online, representing about 60.4% of the top 10 million websites.[1] As of 31 March 2020, the internet was estimated to have around 1.186 billion English speaking users (25.9% of all internet users around the world).[2] In terms of internet penetration, out of the 1.531 billion English speakers estimated for 2021 according to Internet World Stats, 77.5% of them are internet users. The number of English-speaking users has enjoyed a relatively modest growth rate of 742.9% in the last 20 years, compared with Arabic at 9,348%.

2 Technologies and Resources for English

While there has been an increasing interest in developing data and tools for multilingual language processing in the last 20 years, as witnessed by the topics of longstanding shared tasks such as CONLL, nevertheless English continues to be overwhelmingly dominant in every aspect of language processing. This is partially as a result of the dominance of the use and status of English in the digital sphere and as an international language, but also a circular problem related to the availability of existing low-level language processing tools and training data, which provide an easy starting point for further development.

Thousands of corpora are freely available for English. The majority of these are covered by a Creative Commons licence, although they may come with restrictions (e. g., attribution or no commercial use). Some are covered by shared task participation agreements, implying that they are freely available at least to task participants. A number of corpora are released under licences controlled by ELRA and thus only available to ELRA members. The LDC grows by around 30 to 35 new corpora each year, and while these do not all include English, it does mean that new resources with contemporary language use appear with reasonable regularity.

Hundreds of monolingual lexical/conceptual resources are available, most of which are domain-specific, including a few ontologies. It is likely that a huge number of freely available additional resources are available beyond those listed in the main language resource catalogues such as ELRA and LDC. The same is true for bilingual resources that include English. Additionally, a number of multimodal resources exist (where text is one of the forms), mostly concerned with pronunciation.

English is very well-served generally by spelling and grammar-checking tools. Most operating systems have built-in spell-checking tools, for example, aspell and hunspell on Linux. Most programming languages have at least one spell-checking library. Similarly, there are many summarization systems available as open source or commercially, including HuggingFace Transformers. Text-to-speech (TTS) systems are also well supported with a number of open source and commercial models.

There are several major infrastructures or toolkits for language processing available, including GATE, Stanford CoreNLP, Stanford Stanza, NLTK, spaCy, Hugging-

[1] https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/

[2] https://www.internetworldstats.com/stats7.htm

Face Transformers, and OpenNLP, which all contain a variety of processing tools which can be used individually or as a collection. All of these support at least tokenisation, sentence splitting, PoS tagging, and named entity extraction. Some support many more tools such as sentiment analysis, or have specific support for domains such as medicine. Overall, there are thousands of models available, especially for text summarization, translation, TTS and various kinds of classification.

For low-level processing tasks, such as tokenisation, sentence splitting and PoS tagging, there are a few standalone tools and services contained in the ELG platform, but many more are provided as part of standard APIs. In general, tools for tokenisation and sentence splitting for European languages are more or less language-independent. POS tagging is also a reasonably well-solved problem for English.

In terms of Information Extraction, there are dozens of NER systems for English, of which roughly half are domain-specific, with domains/genres including biomedical, Twitter, dendrochronology, environment, chemistry and politics. This is also an area which has seen many ML models released. Tools which fall broadly into the Information Retrieval (IR) category cover a wide range of tasks, including question answering. A number of these are cross-lingual. Many systems enable search in a specified language but can return results in other languages, including English. There are a number of commercial IR engines available, both for generic and specialised tasks. Concerning Machine Translation, there are hundreds of tools, of which a large number contain English as either input or output. The most common pairing (regardless of direction) is English/German.

In terms of LT providers, we have identified 53 major industrial organisations in the UK, including players such as BBC News Labs, the JISC, and Oxford University Press, and 246 research groups or organisations based at 94 different universities. These research groups are split between various faculties and departments, comprising mostly Computer Science and Language departments, but also others such as Medicine, Architecture, Life Sciences and Education, Creative Industries, and Maths. In Ireland there are also extensive LT industry bodies and research centres (e. g., Apple, Accenture, Google, SoapBox Labs, AYLIEN, and CeADAR), whose primary focus is on supporting the English-speaking rather than Irish-speaking population.

3 Recommendations and Next Steps

English is extremely well supported by LT, which is unsurprising given its status in the digital world. Almost every tool and infrastructure or toolkit is first developed to handle English before being applied to other languages. Similarly, an enormous amount of data is available for English. These two factors have a circular effect: due to the amount of data available, training and testing new tools is much easier for English than other languages, and this leads to new models, tools, and resources being developed. The frequency with which English is used for online communication also provides a wealth of data from which to create new corpora, and the availability of a wide range of tools also makes it easier to annotate these with linguistic information. As tools improve, the accuracy and usefulness of pre-annotated corpora also improve, thereby making further tool development easier.

On the one hand, this is an excellent situation for those working on English data, and given the widespread use of English in the digital world, the usefulness of new tools is clear. On the other hand, this can be a double-edged sword for the development of LTs and LRs for other languages. The availability of data, tools and resources for English has fed the enormous success of neural models for developing LT applications, but the lack of data for other languages means that such deep learning models trained on English are not directly applicable. Recently, however, advances have been made in the development of cross-lingual transfer learning in order to build NLP models for a low-resource target language by leveraging labelled data from languages such as English with a high level of resources, or via a staged process whereby training data from English feeds the development of languages with moderate resources, which may have greater similarity to low-resource languages and can feed a further transfer process. Additionally, multilingual transfer settings enable training data in multiple source languages to be leveraged to further boost performance of low-resource languages. On the negative side, almost all languages are inevitably playing "catch-up" compared with English, and as can be seen from our survey, the differences in LTs and LRs available for European languages are striking. It is hard even to grasp a sense of how much is available for English, since resources are so disparate, and the figures reported in the collections of ELG, ELRA and other repositories are only the tip of the iceberg.

References

Ananiadou, Sophia, John McNaught, and Paul Thompson (2012). *The English 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/english.

Maynard, Diana, Joanna Wright, Mark A. Greenwood, and Kalina Bontcheva (2022). *Deliverable D1.11 Report on the English Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-english.pdf.