Chapter 15
Language Report Finnish

Krister Lindén and Wilhelmina Dyster

Abstract During the last ten years, digitalisation has changed the way we interact with the world creating an increasing demand for language-based AI services. In the field of language technology, the Finnish language is still only moderately equipped with products, technologies and resources. The situation has improved in recent years, but still support for automated translation leaves room for ample improvement, as the general support for spoken language is modest in industry applications although some recent research results are encouraging. We take stock of the existing resources for Finnish and try to identify some remaining gaps.

1 The Finnish Language

Finnish is the native language of about 4.9 million people in Finland and the second language of 0.5 million Finns (see Koskenniemi et al. 2012; Lindén and Dyster 2022). Finnish is spoken in several European countries as well as the United States and Australia. Finnish is an official language in the European Union. The Finnish constitutional law and language law define Finnish and Swedish as the national languages of Finland. Moreover, Finnish is an official minority language in Sweden.

The Finnish literary language has a relatively short history. It has been used in religious literature and the church since the 16th century. Laws have been written in Finnish since the 18th century. Until the 19th century, Swedish was used in administration, education and literature, when the foundation of contemporary Finnish was laid and Finnish became a sovereign language in all societal activity.

Dialects are divided into two categories: the Western and the Eastern dialects. The difference is mostly in the pronunciation and word forms (meijän, mänämä in the East, meirän, mennä in the West) and partly in the vocabulary (vasta in the East, vihta in the West). The differences are clear, and speakers from different areas can be identified by their intonation. However, the differences are minor enough to allow speakers of different dialects to understand each other.

Krister Lindén · Wilhelmina Dyster
University of Helsinki, Finland, krister.linden@helsinki.fi, wilhelmina.dyster@helsinki.fi
Finnish is used widely and actively on the internet and social media. Almost all Finnish households (96%) have access to the internet. Traficom, the Finnish Transport and Communications Agency, reported in November 2020 that the total number of registered FI-domains had reached 500,000.

2 Technologies and Resources for Finnish

The development of Finnish language data and tools has progressed steadily over the past 30 years. Since 1995, the Language Bank of Finland\(^1\) and since 2015 CLARIN and FIN-CLARIN have offered a wide variety of text and speech corpora and tools. Today, a large number of fundamental tools and datasets are available for Finnish. Below we present some relevant resources in the different domains of LT.\(^2\)

There are several large monolingual corpora with contemporary textual and spoken language as well as some multilingual corpora. Overall, general domain data seems to be prevalent, e.g., data collected from discussion forums or using web crawls. In addition, news texts, legislative texts and parliamentary speech are well-represented domains. The Language Bank of Finland has the expertise to handle sensitive data, but for example, health domain corpora are still scarce.

The Institute for the Languages of Finland has comprehensive collections of lexical corpora. The Helsinki Term Bank for the Arts and Sciences (HTB)\(^3\) is a multidisciplinary project aiming to gather a permanent terminological database for all fields of research in Finland. The Comprehensive Grammar of Finnish\(^4\) was published in 2004 by the Finnish Literature Society. FinBERT\(^5\) is a version of Google’s deep transfer learning model for Finnish, developed by the TurkuNLP Group. FinBERT is pre-trained with 1 million steps on over 3 billion tokens of Finnish text drawn from news, online discussion, and web crawls. Important software packages are: 1. The Helsinki Finite-State Transducer (HFST)\(^6\) can be used to implement morphological analysers. 2. The Turku Neural Parser Pipeline developed by TurkuNLP\(^7\) is an open source dependency parsing pipeline. 3. The Aalto University Automatic Speech Recognition System (Aalto-ASR)\(^8\) provides functionalities for ASR from audio files and for automatic forced alignment of text and speech. 4. OPUS-MT\(^9\) focuses on the development of free resources and tools for machine translation, with

\(^1\) https://kielipankki.fi
\(^2\) META-SHARE Finland contains additional resources, see https://metashare.csc.fi.
\(^3\) https://tieteentermipankki.fi
\(^4\) https://kaino.kotus.fi/visk/
\(^5\) https://github.com/TurkuNLP/FinBERT
\(^6\) https://hfst.github.io
\(^7\) http://turkunlp.org/Turku-neural-parser-pipeline/
\(^8\) http://urn.fi/urn:nbn:fi:lb-2021082323
\(^9\) https://github.com/Helsinki-NLP/Opus-MT
currently over 1,000 pre-trained neural MT models. 5. Finto AI\(^\text{10}\) is a service for automated subject indexing, which can be used to suggest subjects for texts in Finnish, Swedish and English. 6. Wavelet-based embedding models for speech synthesis for Finnish have been developed at the University of Helsinki.

The Language Bank of Finland supports academic research and provides some support for the industrial use of academic resources which are also available for commercial use. CSC (IT Center for Science) is tasked with providing one of the three EuroHPC supercomputers, LUMI. The whole system is designed with AI, machine learning and data analytics in mind. LUMI’s first pilot phase was concluded by the end of 2021, and LUMI will reach its full capacity in 2022.\(^\text{11}\)

Generally, the Finnish market is extremely active in the AI field. According to the State of AI in Finland report by FAIA (2020), “there are over 1,250 companies that use different AI applications, of which roughly 750 have developed their own technology.” A rapidly growing startup ecosystem boosts AI/LT development.

### 3 Recommendations and Next Steps

In November 2019, VAKE (currently the Climate Fund) published a report (Jauhiainen et al. 2019), specifying the next phase of the language-centric AI development programme and identifying topics in need of intervention. In November 2020, Finland launched an updated national AI strategy. The AI 4.0 Programme promotes the use of AI and other digital technologies in companies, with a special focus on SMEs. In the first interim report,\(^\text{12}\) published in April 2021, the programme presented a vision for the future of the Finnish manufacturing industry, stating that by 2030 the Finnish manufacturing industry will be clean, efficient and digital. As stated in the report, seamless collaboration between high-speed telecommunications networks, cloud computing and AI are central to the digital transformation.

According to the VAKE report, we need the availability and accessibility of components for processing speech with open licences to create prototypes or develop methods into full-scale production versions in the hands of companies. To this end, collaboration between stakeholders is needed: an ecosystem with a forum or a platform where different-level actors can come together to exchange experiences and seek new projects and collaboration opportunities.

Currently, 1. there are some multi-modal resources, but still no advanced discourse processing tools for Finnish; 2. several research projects are working on advanced information retrieval (IR) and data mining for Finnish; 3. the legal situation has become clearer with the General Data Protection Regulation (GDPR), but we are still waiting for Finland to fully implement the Digital Single Market Directive (DSM); 4. we have some specific corpora of high quality, but the commercial sector

\(^{10}\) [https://ai.finto.fi](https://ai.finto.fi)

\(^{11}\) [https://www.lumi-supercomputer.eu](https://www.lumi-supercomputer.eu)

in Finland still needs large, up-to-date resources for product development targeted at everyday users and technologies to collect specialised data sets; 5. work on semantics has still not led to significant applications, but this is explored in the context of advanced research projects on IR and extraction; 6. in speech technology, the recent biggest leaps forward have been made using neural network technology. This has also led to some improvements for the commercial sector offering speech-based services, but speech and video corpora are no longer considered hard to collect with the advent of mobile phones and teleconferencing.

Speech corpora and especially resources for spontaneous speech recognition and various genres of speech synthesis are currently being developed. The need for extensive and varied text materials can to some extent be rectified for research purposes through corpus collections of publicly produced language material when properly considering GDPR and the DSM directive. This will enable the creation of language models. However, we still need a variety of specialised data sets for domain-specific purposes to adapt open-source or proprietary software components. Developing dedicated components from scratch requires giga-scale data sets which may be difficult to compile for small language communities and in specialised domains. This points to a need for a general-purpose language-centric AI which can leverage cross-language and cross-domain resources and benefit from adaptation to local language varieties and specialised domains with small or medium-sized data sets.

References


Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.