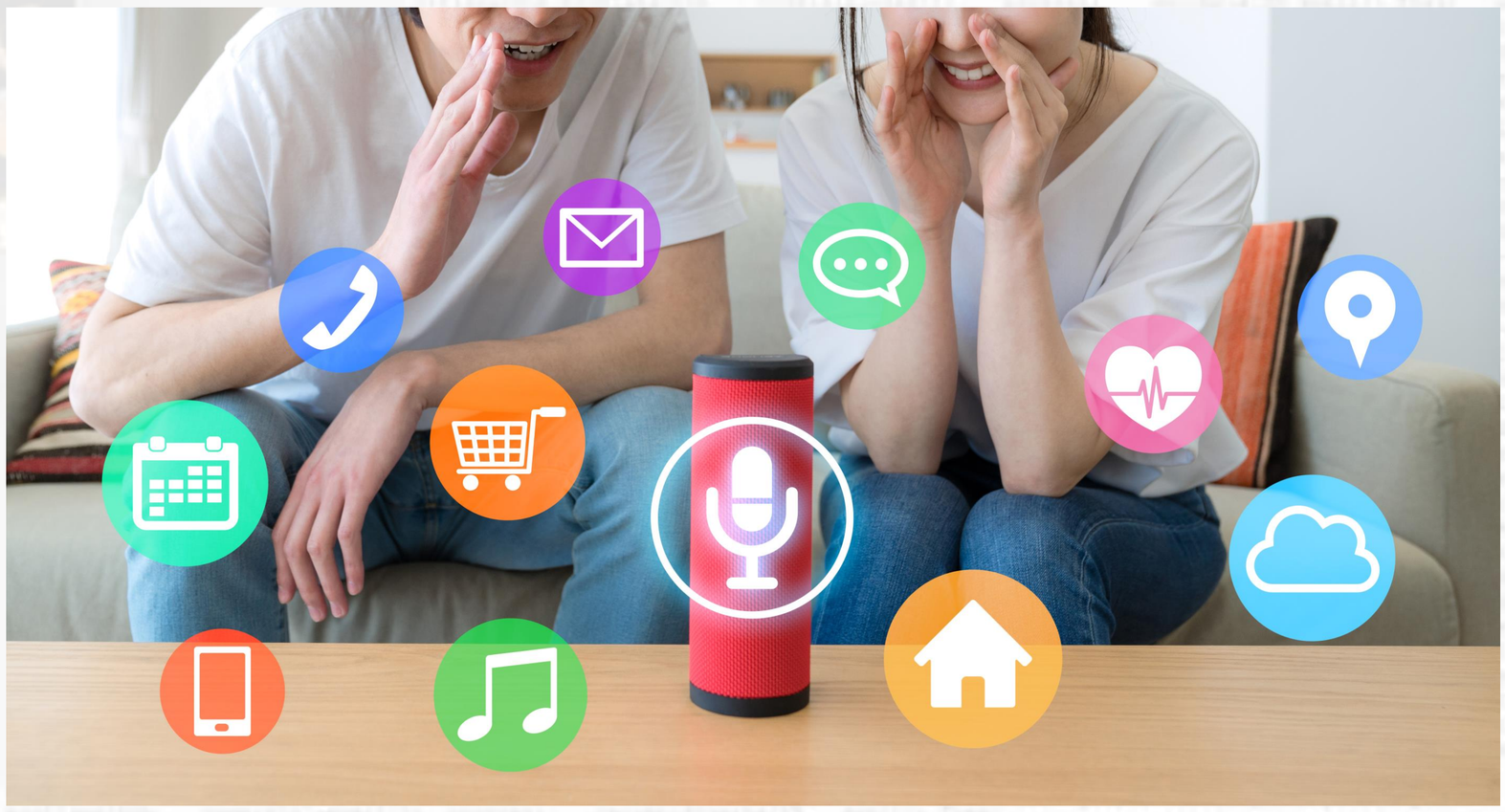
An ambitious move towards ASR that recognises everyone in a country with no spoken standard

Benedicte Haraldstad Frostad
benedicte.frostad@sprakradet.no
The Language Council of Norway

Introduction

The lack of support for one's mother tongue in services and products with integrated automatic speech recognition (ASR) represents a challenge to European and national aims to ensure equal participation in society for all citizens (De Smedt 2012: 41, Directorate-General of the UNESCO 2007). The situation is pressing in Norway, where public and private institutions are increasing integration of ASR. Notably, the courts and The Storting, the Norwegian parliament, are initiating automatic dictation of all court and parliament sessions. This is a challenge in the majority language, Norwegian, which has diversity in both written and spoken forms that is unusual for a national language, and an even greater challenge for minority languages.



Norwegian has two written standards, many dialects and no spoken standard. Speakers are used to linguistic diversity, and dialects are strong identity markers. Changing one's dialect is associated with a loss of identity. The extra costs, need for specialized expertise in spoken Norwegian and lack of suitable lexica and speech data sets complicate the development of ASR products for this language community. This poses a democratic problem as public institutions automatise dictation and integrate ASR as a means for interaction. The Language Council is initiating innovative projects to improve ASR for Norwegian and minority languages in Norway and wishes to exchange ideas and experiences.

Foto: metamorworks / iStockphoto

Language policy and linguistic diversity

Linguistic diversity in Norway
Norwegian is a North Germanic language with approximately 5 mill. speakers, closely related to Swedish and Danish, and it is the majority language in Norway. It has two written and no spoken standard. It has a large number of dialects with significant phonetic, lexical and syntactic variation. (Skjekkeland 1997). There is no spoken variant with a status as an official language. None of the two written standards for Norwegian, Nynorsk (NN) and Bokmål (NB), can be said to correspond to a certain spoken variety. *Målloven* (the Language Act) regulates the use of written standards in the public sector, where each must be used in min. 25% of all text. Both NN and NB allows for significant lexical and inflectional variation. De Smedt and Rosén (1999) demonstrates how a long sentence in Bokmål may be spelled in no less than 165,888 different ways. There is also extensive code-switching between the two. Norwegian dialects have a more prominent role than in other European countries, due to the lack of an official and even a de facto spoken standard (De Smedt et al. 2012: 45), and dialectal variety makes use of ASR challenging in Norway.
In addition, the minority languages Norwegian Sign Language, Kven, North Sami, Lule Sami, and South Sami, Romani and Romanes are protected under Norwegian law.

Norwegian language policy aims to ensure that everyone has the right to a language, to evolve and acquire the majority language, Norwegian, and to evolve, acquire and use their mother tongue, including Sign Language, indigenous languages or national minority languages (The Norwegian Ministry of Culture 2008: 24). Norwegian courts and the parliament are initiating fully automatized dictation for parliament and court sessions. Furthermore, an increasing amount of private and public institutions are communicating with users by means of chatbots, and expect to integrate ASR to these services for increased streamlining and as a means to enable universal design. The Language Council of Norway is responsible for informing institutions about the challenges involved with ASR development and the support of linguistic diversity, as well as to work towards better enabling institutions making use of ASR, as well as developers and researchers to provide products suited for the various Norwegian language communities.

A National Language Bank
Following up on the language policy ambitions stipulated in the parliament white paper *Report no. 35 (2007-2008) to the Storting* (the Norwegian Ministry of Culture 2008), the Norwegian parliament made funds available for a national language bank in 2010, with the aims to collect resources for use in language technology research and development, such as large datasets for text and spoken language, and lexica, available to public as well as private institutions. The National Library is currently responsible for hosting the Language Bank, where resources can be downloaded with no registration necessary by anyone. In 2019, the Parliament decided to make funds available for the development of new resources and it has been decided that the National Library and the Language Council of Norway plan which resources should be developed and made available in co-operation..

Challenges

Scarcity of data
Training an acoustic model and language model for the development of ASR requires sufficient annotated speech data, a pronunciation lexicon (in most cases) and sufficient text data.
The availability of language technology support for Norwegian is extensive considering the size of the language community. However, the lack of a spoken standard and the two written standards makes most products unavailable to a considerable number of speakers. With a few exceptions, products only support one written standard, Bokmål, and speech technology products only support the dialect spoken in the region of the capital, Oslo. Supporting the linguistically diverse Norwegian language community requires more linguistic resources, tailor-made to address linguistic diversity by teams including linguists with expert knowledge of written and spoken or signed variants of Norwegian and Norwegian minority languages.

Costs
Supporting the degree of language variety that is necessary for the development of products and services with integrated ASR that can be used by all Norwegian and minority language speakers is costly. Support for Norwegian requires more resources than for languages such as Dutch and Swedish, where speakers can make use of spoken standards. Norwegian speakers who are not recognised by ASR software in their regional dialect, have no means to standardise their language to be understood. There are few resources available as of yet to support the minority languages. It is therefore important that the government provide resources of high quality, for use by developers and researchers.

Linguistic expertise
The Language Council has learnt through interviews with the developers that computational linguists with sufficient proficiency in the Nynorsk written variant, as well as Norwegian spoken dialects are hard to come by, particularly for developers based outside Norway where most development of Norwegian ASR takes place. Linguistic expertise in the minority languages is significantly scarcer.

Data sharing and information exchange
It is vital that information on products, available resources and linguistic expertise is shared between language communities, developers, institutions making use of products and services with integrated ASR, researchers and the Language Council, to ensure an efficient use of resources and funding. There is currently no efficient infrastructure for the exchange of such information, and the establishment of good networks is a necessary first step towards speech technology that supports Norway's linguistic diversity, and meet language-related political aims.

Current initiatives

The Language Council has initiated two new resources to be developed in co-operation with the Language Bank in 2020 to enable support for dialect diversity and both written standards in Norwegian in ASR development. One is an extension of an already existing pronunciation lexicon with additional transcriptions representing the pronunciation of lexicon items in an additional four dialects. The dialect variant spoken in the Oslo region is already represented in the lexicon. The dialects are carefully selected in co-operation with the University of Oslo, to represent all five dialect areas, and as much lexical and phonetical variation as possible. All transcriptions will be tagged for dialect, and developers and researchers can select the ones they want to include. The pronunciation lexicon can be used for both ASR and speech synthesis (TTS). The other is a speech database covering all five aforementioned dialects, tailor-made for digital assistants, notably automotive and mobile assistants. The National Library has taken the initiative to create a large speech database consisting of annotated parliament sessions. A wide variety of dialects are used in parliament, and the annotation will be available in both written forms. The Language Council is reaching out to stakeholders to plan projects aimed at ASR support for Norwegian minority languages.

Towards ASR that recognises all

Close co-operation with developers, researchers and language communities
To ensure that funding for the development of resources to be used for research and development of language technology is used in an as efficient way as possible, the Language Council needs close communications with resources and institutions planning on using language technology resources. It is important to put emphasis on products that are under development or will be developed and used in the near future, rather than products that may not be developed anytime soon. Thanks to efficient communication with developers and public institutions, The Language Council learned that it is currently vital to create speech datasets for automotive and mobile assistants and parliament sessions. However, we can do better. An efficient infrastructure for information exchange needs to be established to provide the resources and advise needed to researchers, developers, public institutions and, above all, the users, the members of the many Norwegian speech communities themselves, that need access to new technology.

Sign language recognition
A project needs to be initiated and funded where gesture recognition researchers and developers collaborate with Norwegian sign language linguists on exploring the possibility for Norwegian sign language recognition.

Alternative methods for ASR developments for severely under-resourced languages
The Language Council and developers are collaborating on exploring the possibilities for developing ASR for the Kven language, which could possibly be combined with the closely related language *Meänkieli, spoken in Sweden*. Based on comparable projects, where e.g. ASR for Afrikaans was successfully developed with a system partly trained on Dutch, possibilities to use an acoustic model based on Finnish, a closely related language, is explored. Unfortunately, the Sami languages do not have well-resourced languages that are close enough in linguistic proximity for this to be an opportunity.

Continued development of language resources
Norwegian and all minority languages in Norway are severely under-resourced. As of today, no pronunciation lexicons exist for any minority language, and there is only one for 350,000 words in Norwegian Nynorsk. Speech databases for Norwegian lack sufficient dialectal coverage and coverage for age and gender. The Language Council is reaching out to all stakeholders to plan the development of quality-assured resources suited to develop ASR technology that works for all. Some projects are under way, but many more are needed.

References
De Smedt, Koenraad, Gunn Inger Lyse, Anje Müller Gjesdal and Gyri S. Losnegaard. 2012. *The Norwegian Language in the Digital Age – Norsk i den digitale tidalderen*, METANET Whitepaper, Berlin: Springer.
De Smedt, Koenraad and Victoria Rosén. 1999. «Automatic proofreading for Norwegian: The challenges of lexical and grammatical variation» in *Proceedings of NOVALIDA 1999*.
The Norwegian Ministry of Culture. 2008. *Report no. 35 to the Storting (2007-2008): Mål og Meining – Ein heilskapleg norsk språkpolitikk*, Oslo: Akademika AS.
Skjekkeland, Martin. 1997. *Dei norske dialektane – Tradisjonelle særdrag i jamføring med skriftmåla*, Kristiansand: Høyskoleforlaget.