

Towards ASR that Supports Linguistic Diversity in Norway

Benedicte Haraldstad Frostad, Verena Schall, and Sonja Myhre Holten

The Language Council of Norway

Observatoriegata 1 B, 0254 Oslo, Norway

{benedicte.frostad, verena.schall, sonja.myhre.holten}@sprakradet.no

Abstract

Norway's majority language, Norwegian, has two written standards, many dialects and no spoken standard. Norway also officially recognises some national minority languages. The extra costs, need for linguistic expertise and lack of suitable lexica and speech data sets complicate the development of ASR products for all these Norwegian language communities. 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.

Keywords: linguistic diversity, dialects, written standards, spoken standards, automatic speech recognition, Norwegian, minority language, Norwegian sign language, indigenous languages, Kven, Sami, Romanes, Romani

Résumé

Majoritetsspråket i Norge, norsk, har to offisielle skriftnormer, mange dialekter og ingen offisiell uttalenorm. Norge anerkjenner også noen nasjonale minoritetsspråk. De ekstra kostnadene, behovet for lingvistisk ekspertise og mangelen på passende leksika og taledatasett vanskeliggjør utvikling av talegjenkjenningsprodukter for alle disse språksamfunnene. Det er et demokratisk problem ettersom offentlige institusjoner innfører automatiske dikteringsverktøy og integrerer talegjenkjenning i sine kommunikasjonskanaler. Språkrådet har tatt initiativet til nyskapende prosjekter skreddersydd for å øke kvaliteten på talegjenkjenning for norsk og minoritetsspråk i Norge og ønsker å utveksle idéer og erfaringer.

1. 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.

Norway has a long history of Norwegianisation policy directed towards sign language users, indigenous and national minorities. This policy has been carried out differently towards each community, thus, the vitality of Norwegian minority languages and their language communities varies, as do the resources available for each language. This is particularly noticeable in the field of language technology, where research and development rely heavily on language resources of high quality and linguistic expertise.

2. Linguistic diversity in Norway

2.1 Norwegian

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.

2.1.1 Written Norwegian

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.

2.1.2 Spoken Norwegian

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.

2.2 Norwegian Sign Language

Norwegian Sign Language (NSL) is among the largest minority languages in the country, with an estimated 16,500 speakers (the Norwegian Ministry of Culture 2008). Its status is nonetheless precarious. During the last decade three out of four national deaf schools have closed down, as a new policy to integrate children with hearing loss in mainstream schools was introduced. Negative attitudes to sign language and the loss of communities due to inclusion in mainstream schools, have led to a less vital status. Political aims for NSL are two-fold. The first is the so called democratic perspective (as stipulated in the Convention on the Rights of Persons with Disabilities (CRPD)) where key words are, access (to the mainstream society), and facilitation. Access and facilitation, because signers who are deaf or hard of hearing are more or less relying on sign language, as this is the only language

modality in which they can interact freely and effortlessly with their interlocutors. The second is the language policy perspective. NSL has a value in itself. It is a culture and an identity marker, as much for hearing people as deaf/hard of hearing, and in addition part of our cultural heritage. Today, the NSL operates with two geographical variants, one in the north part of the country, and one in the south.

NSL has been legally recognised in Norway through the Education Act as a minority language since 1997. Norway have also ratified CRPD, and has thus committed to promote and recognise the use of sign language, and to implement universal design. NSL was recognised as a part of the Norwegian language diversity and a part of the cultural heritage of Norway in The Norwegian Ministry of Culture (2008) and this right was affirmed in the Storting in 2009.

2.3 The Sami languages

The Sami languages belong to the Uralic language family. There are mainly three Sami languages in use in Norway: North, Lule, and South Sami, which to some extent are mutually intelligible. The North Sami Language has the largest language community. In comparison, Lule Sami and South Sami are far more endangered with fewer language users and fewer resources.

2.4 Kven

The Kven language belongs to the Finno-Ugric group of the Uralic language family. It is heavily influenced by Norwegian and the Saami languages, and closely related to Meänkieli, a national minority language in Sweden. Due to suppressive national language policies, the Kven ethnic minority outnumbers the language community. There are no official numbers, but one presumes that the former counts about 50,000-60,000 members, whereas estimates for the number of speakers range between 1500 and 10,000 (Schall 2017).

2.5 Romani

Romani (also known as Scandoromani) is spoken by the Romani ethnic minorities (some members of the community prefers to be referred to as Travellers) in Sweden and Norway. It has a North Germanic grammatical structure and a heavy lexical influence by Romanes.

2.6 Romanes

Romanes is spoken by the Roma ethnic minority in Norway, and is internationally often referred to as Romani. It is an Indo-Aryan language with many varieties all over Europe. The Norwegian varieties belong to the Northern Vlach-group. Bilingualism between several variants is fairly common, and many speakers speak two or more variants, especially Lovara and Kalderaš. It is estimated that the Roma community in Norway count approximately 700 individuals, and the continued marginalisation of the community has a significant impact on the language community. Studies indicate a lack in children's formal education and a high illiteracy rate among adults (Hagatun 2019). Romanes is mainly used as a spoken language, and there are few written sources of the language in Norway. Most Roma children start school mainly Romanes speaking, an indication of a healthy spoken language and conscious language planning in families and within the

community. Unfortunately, authorities have shown little initiative to protect the language. Developing language infrastructure is essential and will be valuable for educational practice, but as the varieties of Romanes spoken in Norway are still poorly documented, this is a challenging task.

2.7 Yiddish

Yiddish is a West Germanic language with considerable influence from Hebrew, Aramaic and Slavonic languages, with a long history as a minority language in Norway. It is not recognised in the European Charter for Regional or Minority Languages (ECRML) The Language community, however, has certain language-related rights through the Framework Convention for the Protection of National Minorities, but unfortunately this framework is less specific than ECRML. Yiddish was a vital language pre-WWII, but is now near-extinct in Norway.

3. Language policy ambitions

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). To meet these goals, it is important that language use in digital interactions are not left out of the picture.

3.1 Language policy and ASR

Norwegian courts and the parliament are initiating fully automated 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 in 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, and to better enable developers and researchers to provide products suited for the various Norwegian language communities.

3.1.1 The Norwegian 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.

4. Linguistic diversity and ASR

4.1 Challenges

4.1.1 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.

4.1.2 Costs

Supporting the degree of language variety that is necessary for the development 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 in order 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.

4.1.3 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.

4.1.4 Data sharing and information exchange

To ensure an efficient use of resources and funding, 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. 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.

4.2 Existing resources

4.2.1 Resources in the Language Bank

The Language Bank contains several resources for

language technology in Norwegian, developed as part of public research projects or by private companies. As of now, it contains few resources suited to address spoken and written diversity support needs in ASR development and no relevant resources for minority languages.

4.2.2 Resources in CLARINO

Resources available in the Norwegian CLARIN database are mostly suited for research in language technology and linguistics. CLARINO data mostly covers Norwegian, but will contain a large dataset for NSL in the near future.

4.2.3 Resources in Giellatekno and other institutions

Resources for the Saami languages and Kven are largely managed in cooperation by Divvun, the Kven Institute and Giellatekno, an open source repository. Some text processing tools and text databases exist for Kven. There are a few more for the Saami languages, due to a TTS project. There are pronunciation lexica, text and speech databases and various other tools for Sami languages. There are some lexical resources for Norwegian Romani made available on <https://app.uio.no/hf/nro/index.php?link=contact> from a Ph. D. project.

4.3 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 four additional 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.

NSL is a visual-gestural language and like all signed languages it lacks a written form and is primarily a face-to-face language. Thus, in order to create accessible and open documentation of NSL, a digital (online) platform able to host a large collection of video-recordings of the language is needed. Current initiatives in Norway, such as the CLARINO infrastructure project, show potential, but financial and technological support is needed to ensure that documentation of NSL can be archived and preserved over the long-term.

4.4 Towards ASR that supports linguistic diversity

4.4.1 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 must be established to provide the resources and advice 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.

4.4.2 Sign language recognition

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

4.4.3 ASR development for under-resourced languages

The Language Council and developers are collaborating on exploring the possibilities to develop 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, the possibility of using 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.

4.4.4 Universal design

It is vital that universal design is considered at all times when planning and developing new resources. Norway's Equality and Anti-Discrimination Act stipulates that all public and private undertakings focused on the general public have a duty to ensure that their general functions have a universal design, including all ICT solutions. Considering the limited availability of funds, developers and linguistic expertise, it is important that all resources that are developed are well planned and contribute to the development of technology that satisfies legal requirements.

4.4.5 Development and dissemination of linguistic 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.

5. Conclusion

More funding and governmental initiative are needed to secure ASR that is accessible to speakers of Norwegian and Norwegian minority languages. Long-term planning and stable, large projects are needed to provide resources of the quality that is needed for ASR and sign language recognition that supports Norway's linguistic diversity. Careful planning of resources requires an infrastructure that allows for efficient information and data exchange between representatives of the language communities, the Language Council, research communities, developers, public institutions, and institutions developing resources and making them available (through interfaces such as the Language Bank). It is particularly beneficial for sign language recognition and ASR development for minority languages that this infrastructure also includes international collaborators. Governmental institutions need to make informed decisions when applying ASR, and take linguistic diversity into consideration. the Language Council has addressed this issue when reaching out to and meeting relevant institutions, and this needs to be followed up.

6. Bibliographical 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 tidaldere*, 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*.
- Directorate-General of the UNESCO. 2007. *Intersectional Mid-term Strategy on Languages and Multilingualism*, http://unesdoc.unesco.org/images/0015/001503/150335_e.pdf
- Hagatun, K. 2019. “They assume that I don't really want education for my children”: Roma mothers' experiences with the Norwegian educational system” in *HERJ Hungarian Educational Research Journal. Special issue.*, 9(1), 9-21. doi:10.1556/063.9.2019.1.2
- The Norwegian Ministry of Culture. 2008. *Report no. 35 to the Storting (2007-2008): Mål og Mening – Ein heilskapleg norsk språkpolitikk*, Oslo: Akademika AS.
- Schall, V. 2017. «Språk, identitet og minoritetspolitikk [Language, identity and minority policy]» in N. Brandal, C.A. Døving, & I. Thorson Plesner (Eds.): *Nasjonale minoriteter og urfolk i norsk politikk fra 1900 til 2016*. Oslo: Cappelen Damm Akademisk.
- Skjekkeland, Martin. 1997. *Dei norske dialektane – Tradisjonelle særdrag i jamføring med skriftmåla*, Kristiansand: Høyskoleforlaget.