DigiSami and Digital Natives: Interaction Technology for the North Sami Language

Kristiina Jokinen, Katri Hiovain, Niklas Laxström, Ilona Rauhala
and Graham Wilcock

Abstract The paper discusses revitalisation of endangered languages and proposes that interactive talking robot applications can make an important contribution. The goal of the DigiSami project is to improve digital visibility and viability of the North Sami language, by developing technological tools and resources that can be used for speech and language processing and for experimenting with interactive applications. We describe the first steps in the development of SamiTalk, a Sami-speaking robot application which will allow North Sami speakers to interact with digital resources using their own language.

1 Introduction

In this paper we discuss revitalisation of endangered languages and present our work in the DigiSami project, whose goal is to apply language and speech technologies to support the North Sami language community. We propose that an interactive talking robot application can make an important contribution towards this goal.

The paper is structured as follows. Section 2 discusses revitalisation and presents the project goals, to improve digital visibility and viability of the Sami language. We provide an overview of the Sami languages and of existing tools and resources for North Sami that can be used for speech and language processing. Section 3 describes the DigiSami Corpus of spoken North Sami and gives a preliminary analysis of the conversations in terms of laughter, speech properties, and change in the function of adjective forms. Section 4 describes first steps towards developing SamiTalk, a spoken dialogue system for Sami-speaking humanoid robots. Based on previous work on the WikiTalk system, SamiTalk will provide spoken information access from Sami Wikipedia. The DigiSami Corpus is being used to develop the speech components and to model dialogue. Section 5 presents conclusions and future work.

University of Helsinki, Helsinki, Finland e-mail: first.last@helsinki.fi

2 Revitalisation of endangered languages

The digital revolution of our era has made a dramatic impact on nearly all aspects of society. Global economics, technology and politics produce interdependence of countries world-wide, while everyday life is drastically changed by a media-rich environment where communication technology brings people speaking different languages together in new ways. New genres of discourse emerge through social media platforms and applications, collaboratively edited content, and user-generated online materials. The role of language in these novel situations is prominent, since language is the vehicle that manifests these changes and which must also adapt itself to these changes. The new paradigms thus affect language use, and one result of this is the endangerment of minority or lesser-used languages: these language communities are the most sensitive to outside forces and therefore most affected by the new paradigms in communication technology.

A language can survive only if it is in active use in a variety of interactive contexts, including new media social networks, business and commerce, live literature/blogs, etc. In other words, in order to make a language viable in the web and digital world, it needs to have a function that is performed digitally. For this purpose, it is important to have tools and applications which support the language use in the new communication paradigms, and in order to develop these, it is necessary to have quality resources and corpora.

However, cutting-edge enabler technologies of language processing applications are typically available only for widely-spoken languages (so called ”comfort zone” languages), while smaller communities are often left to their own resources, or they need to translate or localize information that is unavailable in their native language. The needs of less-resourced languages are to be specifically considered, in order to reduce the unbalanced situation among languages, see [21].

2.1 The DigiSami project

The DigiSami project, within its larger framework, sets out to investigate how modern language technology and corpus-based linguistic research can contribute to facing the above challenges. As described in [10], the project aims to

- collect data from dialogue-related genres and annotate it on a range of levels from grammatical to discourse phenomena,
- experiment with language technology applications that can strengthen the user's interest in using the language in various interactive contexts, focussing in particular on the SamiTalk application,
- alleviate barriers in accessing information from user-generated content, supporting community-based generation of translated material on the web, based partially on existing language resources and technology.

In this paper we focus on the first two goals of the project and describe how the SamiTalk robot application can be used to support revitalisation of the North Sami language. The robot application is chosen because it is an interface to collaboratively edited Wikipedia information, and as a novel application it is expected to increase the visibility of the languages as well as interest in them. In particular, it is expected that young people may become more interested in using the language, which is an important and effective strategy for language revitalization in general.

2.2 The Sami languages

We have chosen North Sami as our target language, since it is the largest of the Sami languages and generally used as a lingua franca among the Sami people. The areas where Sami languages are (or were) spoken is shown on the map in Figure 1.

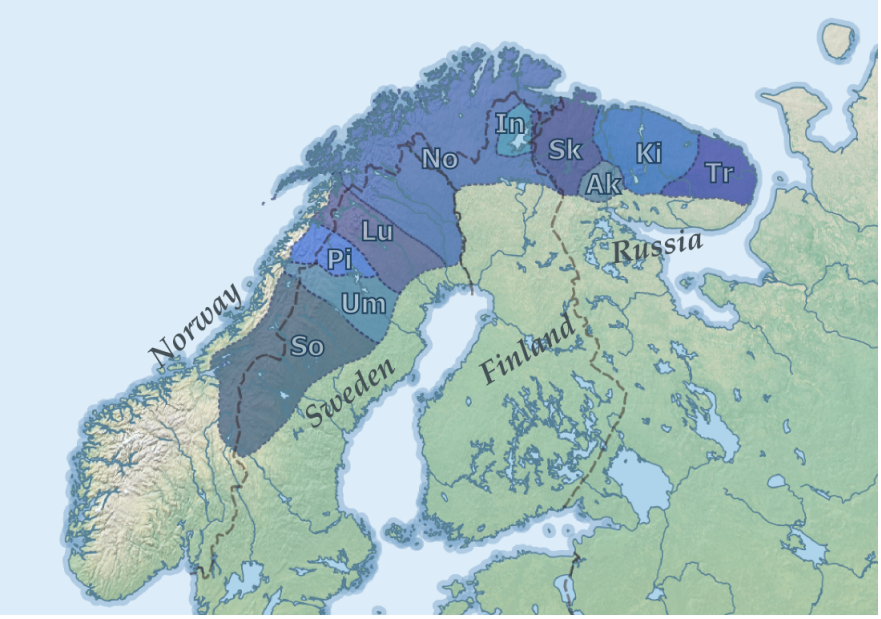

Fig. 1 The Sami language areas at the beginning of the 20th century. Source: [19].
Abbreviations: So - South Sami, Um - Ume Sami, Pi - Pite Sami, Lu - Lule Sami, No - North Sami, In - Inari Sami, Sk - Skolt Sami, Ak - Akkala Sami, Ki - Kildin Sami, Tr - Ter Sami.

The Sami languages are close relatives and the distinction between a dialect and a language is sometimes vague. The differences mainly concern morphophonetic variation whereas syntactic changes are fairly small [18]. The closer the languages are geographically, the more easily speakers understand each other, but also the

majority language affects understandability since it is reflected in differences in both vocabulary and pronunciation as described in Section 3.2.

The Sami languages belong to the Uralic language family which includes Finnish, Estonian, and Hungarian. They are divided into eastern and western groups, shown in Table 1. Both language groups are represented in Finland where North Sami, Inari Sami and Skolt Sami are spoken. The last speaker of Akkala Sami died in 2003.

Table 1 The Sami languages, their region and estimated number of speakers in 2012 [9].

Western Sami languages	**Eastern Sami languages**
South Sami (500 speakers) Norway, Sweden	Inari Sami (350 speakers) Finland
Ume Sami (5 speakers) Sweden	Skolt Sami (300 speakers) Finland, Russia
Pite Sami (40 speakers) Sweden	Kildin Sami (700 speakers) Russia
Lule Sami (1000 speakers) Norway	Ter Sami (2 speakers) Russia
North Sami (30,000 spkrs) Norway, Sweden, Finland	Akkala Sami (0 speakers) Russia

The Sami languages use the Latin alphabet with various diacritics to represent phonological differences, except for Kildin Sami which is spoken only in Russia and written using the Cyrillic alphabet. Although there have been written Sami texts for more than 200 years, their orthographies were irregular until the 1970's and 1980's when national educational reforms gave a legal basis for Sami language education, and even today Sami speakers who did not learn to write in Sami at school may still feel uncertain of the written language conventions.

In the case of North Sami, the Sami Council decided on its modern orthography already in 1978 [13]. However, because it is spoken in Norway and Finland where the different majority languages have different orthographic conventions, there are still irregularities in the way it is written. It is not uncommon to find "mistakes" in crowd-sourced texts such as Wikipedia articles, and this makes their automatic processing a challenge. The normative work continues today in collaboration with representatives from all the relevant countries.

Nowadays the Sami language (the term used to refer collectively to all the Sami language variations and to emphasise the Sami as a nation) is officially recognized in Norway, Finland, and Sweden. The respective Language Acts (1992 in Finland and Norway, 2000 in Sweden) guarantee the official status of the Sami language and the right of the Sami to use the Sami languages in all official encounters. Moreover, as a result of the national education reform in Finland, the Basic Education Act 1998 entitles Sami children who live in the Sami Homeland area and speak the Sami language to receive the main part of their basic education in the Sami language.

2.3 Existing North Sami language resources

Technical development has played a significant role in revitalizing and modernizing the Sami languages. Nowadays several language tools are available for speakers and

language learners to check spelling, look for suitable words, and learn the language. Special keyboards adapted to the Sami orthographies are available for computers and cell-phones [1], so as to facilitate Sami language text input, which can be challenging due to various diacritics and their different encodings.

North Sami enjoys a relatively favourable situation as it has various tools for automatic language analysis. They have been developed at University of Tromsø and are available from Giellatekno [7]. For instance, the spell-checker *Divvun* supports the writing of North Sami, and has an important role in taking the written language norms into use. Morphological and syntactic parsers for text corpora are also available, as well as a translation tool from North Sami to Norwegian Bokmål.

A new approach to Sami morphology is taken in [8] which studies how to use an Active Learning approach to morphological segmentation. Since high-quality morphological analyzers require a significant amount of expert labour, data-driven approaches may provide sufficient quality for many applications. [8] describes how the semi-supervised Morfessor FlatCat method is used to create a statistical model for morphological segmentation for a large unannotated corpus, with a small amount of annotated word forms which are selected using an active learning approach.

Revitalization of a language also means gaining new speakers via language learning. For this there are digital dictionaries and also a language learning website *Oahpa!* [3]. It offers different ways to learn and test language skills, with exercises for testing and practising morphology, vocabulary and syntax. The digital dictionaries [6] include more updated vocabulary than printed ones, and it is possible to search for a word, both from a majority language to Sami and vice versa.

Various spoken and text corpora are also available in North Sami. These are at the Sami culture archive [5] in the Giellagas institute at University of Oulu. The collection of audio and video material as well as photographs and written documents supports research infrastructure, and documentation of the Sami culture. There are no conversational corpora but there are spoken corpora of interviews and official texts, such as the corpus of Yle Sápmi radio programs [4].

The DigiSami project has collected spoken data in North Sami (see Section 3). The corpus has been transcribed and translated into Finnish, and is unique in that it contains natural conversations between a group of people.

2.4 Existing North Sami speech technology

Current speech technology applications for the Sami languages still require development and are not yet commonly in use. A big challenge is the limited speech corpora available. Some speech/voice data is available for North Sami under licence of the Sami Parliament of Norway, but speech corpora for other Sami languages remain in limited use at the moment. However, some advance has already been made in the development of a speech synthesizer and a speech recognizer for North Sami.

A North Sami speech synthesizer *North Sami Infovox 4* (Windows) or *North Sami iVox* (OS X) was developed by Divvun and the Norwegian Sami Parliament,

in cooperation with the voice and speech technology company Acapela [2]. It was released in May 2015. The system has both a female and a male voice, and they can be adapted to the user's needs. North Sami speech synthesis has also been studied in the Simple4All project [20] which focuses on creating methods that enable speech synthesis systems to be built by little or no supervised learning from the data.

A North Sami speech recognizer has been developed in the DigiSami project in collaboration with our partners at Aalto University. [16] describes building an automatic speech recognizer for North Sami and discusses its further development. This is a notable work since to the best of our knowledge, this is the first and only speech recognizer for any of the Sami languages today.

3 The DigiSami Corpus

The DigiSami Corpus of spoken North Sami consists of both read speech data (altogether 257 minutes of annotated data) and conversations of two or three persons (altogether 195 minutes of annotated data). It was collected in Enontekiö, Utsjoki, Inari and Ivalo in Finland, and in Kautokeino and Karasjok in Norway, in the areas traditionally inhabited by the Sami (see more details in [11, 10]).

The annotation of the corpus was done with Praat and consists of 5 time-aligned tiers: a phonological/phonetic transcription, the words, the sentence in orthographic form, a Finnish translation, and remarks on things like dialectal variation. It was performed by two non-native annotators. The conversations contain some unclear speech, some unknown names, as well as some insider jokes that were difficult to understand and translate. The corpus collected from Norway includes Norwegian words that also were challenging.

The speakers are all native speakers of North Sami, and their ages vary between 16 and 65 years (see [11] for details). The style of some conversations is familiar when the participants knew each other beforehand and referred for instance to things they had been talking about earlier. The styles of the conversations differ depending on the age of the speakers and their hierarchical difference. The conversations between a pupil and his/her teacher are more like interviews than conversations, and the topics stick to things that one could write a Wikipedia article about (such as Sami language, Sami costume, music, reindeer herding, and snowmobiles).

The participants discussed freely about their own interests but also about the Wikipedia articles they were to write. The conversations between young students concern more their everyday life, and the topics are the next vacation, driving school, and cars. The conversations between two adult men who have known each other for a long time concern translation between Sami and other languages, and the technological tools that have been made to help writing North Sami more correctly.

The conversational corpus is unique among the Sami language corpora because it contains natural conversations between the participants, and because it is multimodal, i.e. conversations are both recorded and videotaped. This allows us to study the language as it is, not as it somehow should be. The spoken corpus helps to cre-

ate a model of colloquial language for the use of speech technology applications, differing from the formal grammars and dictionaries.

3.1 Preliminary analysis: engagement and interaction

The participants' engagement in the conversation and mutual bonding can be measured using multimodal and non-verbal cues, such as the amount of laughing or chuckling, and overlapping speech. For the purposes of measuring engagement, we annotated the data with these features on the remarks tier in Praat to see how they act as a part of conversations and what they can tell about the interaction situation. The basic statistics are shown in Table 2.

Table 2 Laughter, overlapping speech and use of words from majority language

Region and conversation code	Informant code	Laughter	Overlapping speech	Use of word from majority language
Karasjok, Norway 01_S	S-1	9	15	3
	S-2	25	3	12
	S-3	9	4	24
Karasjok, Norway 02_S	V-1	75	1	5
	V-2	34	–	13
	V-3	63	1	3
Ivalo, Finland 03_V	V-2	7	1	–
	V-3	6	–	–
Ivalo, Finland 04_S	S-1	0	–	–
	S-2	6	–	–
	S-3	1	–	–
Ivalo, Finland 05_TP	TP-2	21	–	2
	TP-3	34	–	2
Utsjoki, Finland 06_PS	PS-1	5	–	2
	PS-3	4	–	–
Utsjoki, Finland 07_SX	SX-1	3	–	2
	SX-X	1	–	–
Utsjoki, Finland 08_VV	VV-Vih	15	–	–
	VV-Vio	19	–	–

The analysis of different types of laughter shows some connections to how well informants know each other, how nervous they are, and what kind of relationship they have with each other. For example, in the conversation in which the informants laugh and chuckle the most, they seem very nervous, and their conversation topics change very fast and have "silences" in them. The lack of laughter may indicate a rather formal conversation and asymmetrical relationship between speakers, such as teacher-pupil conversation. In fluent conversations between young people that know each other and with no impression of nervousness, laughter occurred only when telling jokes or funny stories. These observations will be substantiated with deeper statistical analysis, and models for joking, laughing and generally positive

attitude will be explored further so as to enable appropriate models be implemented in the SamiTalk application.

Turn-taking was usually performed smoothly and the overall amount of overlapping speech was low. There was remarkable overlapping speech only in one fluent conversation with three young female participants from Norway (see Table 2).

3.2 Preliminary analysis: influence of majority language

There are also differences in the influence of the majority language on vocabulary. The use of Norwegian words in Sami in Karasjok is much more common than the use of Finnish words in Ivalo or Utsjoki as shown in Table 2.

The corpus indicates differences in spoken Sami between Norway and Finland, although dialectally the regions of Ivalo, Utsjoki and Karasjok represent the same eastern Finnmark dialect of North Sami. The differences can be heard in different pronunciation of phonemes, such as /r/. In our data from Karasjok in Norway, /r/ has several allophones, occurring as [ø], [r], [ɪ] and [ɾ], while in our data from Finland it occurs only as [r] with no allophonic variation. The IPA transcriptions and spectrogram pictures from the speech of two young girls in Figures 2 and 3 demonstrate this difference (the example word is *fertet* 'have to').

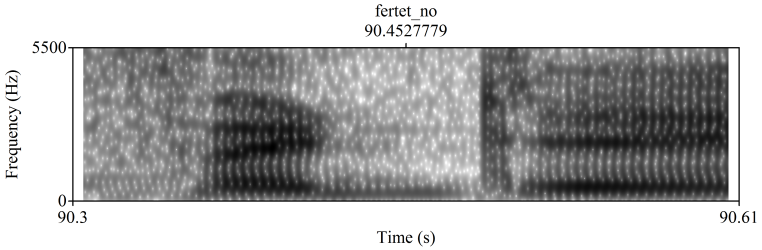

Fig. 2 Sami spoken in Norway: [fəɪdːeːɪ]

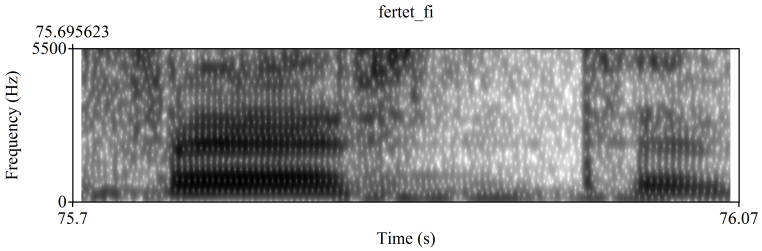

Fig. 3 Sami spoken in Finland: [feːrhtːe]

3.3 Preliminary analysis: adjectives in spoken language

The corpus data shows that use of adjectives in spoken language differs from what is presented in grammars and dictionaries. This indicates that the adjective system is changing. In North Sami the adjective attributive form tends to be different from the nominal predicate form, e.g. *fiinnis* (pred.) *fiinna* (attr.) 'fine, nice'. The attributive marking system is complex [19], and the preliminary findings indicate that the system is changing and the main factor of the change is frequency. The frequency of an adjective being in a certain position affects the preservation of form. For example, the adjective *ođas* 'new' (altogether 8 occurrences in the corpus) occurs mostly in attributive position (6 times). The attributive form is *ođđa*, and the base form *ođas* is also the nominal predicate form. It occurs only twice as nominal predicate, and in these cases one speaker hesitates looking for the right form, while the other speaker uses the attributive form *ođđa* as nominal predicate. Interestingly, the age of a speaker does not affect the use of forms; both older and younger speakers hesitate or make mistakes in the use of adjective forms. Further analysis of the changing mechanisms of the adjective system is essential in order to provide realistic data of the spoken language for developing the SamiTalk application.

4 Towards SamiTalk: a Sami-speaking robot application

Often endangered languages face a gradual language death by assimilation. The ability to use one's own language with new technology, in the modern world, is almost a necessity to prevent gradual language death. In these cases the motivation to continue using the endangered language is among the most important factors.

Robot applications are leaving the research laboratories and reaching the general population, both in homes and outside home. For example, WikiTalk [23, 12] is a multilingual spoken dialogue system that runs on a Nao robot. The user and the robot have a dialogue in which the robot talks fluently about an unlimited range of topics using information from Wikipedia.

Localisation [15] of robot applications to an endangered language benefits the language in multiple ways by providing motivation to use the language. Localised robot applications can have a favourable effect on the prestige of the language, by showing that efforts have been made to support the language on the new technology. At home, applications such as WikiTalk can prevent bottom-to-top language death by encouraging use of the language in the family. Of course, with a multilingual application there is a risk that the users will just switch it to some other language, or even more likely, will not switch it to the local language from the default language, because they do not know how, or do not even know that their language is available. For this reason, language selection [15] for robot applications is an important topic to study further.

New technology also offers opportunities for language revitalisation including existing or proposed Wikipedias for endangered languages. Wikipedias can bring

together speakers of endangered languages even if they are not physically close to each other, can give motivation by Wikipedia's mission to provide free knowledge to everyone [22] and can foster collaboration between speakers and scholars studying the endangered language. There is a risk, however, that a Wikipedia started by scholars may fail to attract native speakers, for example if the native speakers are unable to access Wikipedia via desktop computers or mobile devices.

The DigiSami project is working towards the creation of SamiTalk, an interactive robot application in the North Sami language, as part of its support for language revitalisation using speech and language technologies. The SamiTalk application will be based on existing WikiTalk technology [12] and will provide access to Sami Wikipedia information via a dialogue in North Sami with a humanoid robot. This work is described in more detail by [24].

Up to now, the available Wikipedias in the languages supported by WikiTalk have been very large. English Wikipedia has almost 5 million articles, Japanese Wikipedia has almost 1 million articles and Finnish Wikipedia has over 350,000 articles. In these languages, WikiTalk can talk about almost any topic the user is interested in. By contrast Sami Wikipedia is much smaller, with about 7000 articles. This means that there are many topics that SamiTalk will not be able to talk about using existing methods. To address this problem, the DigiSami project is supporting initiatives to encourage the North Sami community to create new articles in Sami Wikipedia [11], and is investigating methods for on-line translation of Wikipedia articles [14] into under-resourced languages.

5 Conclusions and Future Work

The DigiSami project has collected a conversational spoken language corpus for an endangered language, North Sami. The corpus has been transcribed and annotated, and preliminary analysis already shows interesting properties concerning laughing and joking, as well as turn-taking and overlapping speech. We are taking the first steps in developing SamiTalk, an interactive robot application for North Sami. This work is based on the existing WikiTalk system, but SamiTalk requires new speech components for North Sami and new spoken language corpora. We are developing speech technology for the language with our collaborators. Future work will deal with integration of the speech technology on the humanoid robot as well as deeper analysis of the conversational corpus.

References

1. Divvun, University of Tromsø: Keyboards (2015). URL http://divvun.no/keyboards/index.html. [Online; accessed 2-December-2015]
2. Divvun, University of Tromsø: Text-to-speech (2015). URL http://divvun.no/en/tale/tale.html. [Online; accessed 2-December-2015]
3. Divvun, University of Tromsø: Welcome to the OAHPA! portal (2015). URL http://oahpa.no. [Online; accessed 2-December-2015]
4. Giellagas Institute, University of Oulu: A. Äänitteet (Recordings) (2015). URL http://www.oulu.fi/giellagasinstituutti/aanitteet. [Online; accessed 2-December-2015]
5. Giellagas Institute, University of Oulu: The Saami Culture Archive of University of Oulu (2015). URL http://www.oulu.fi/giellagasinstitute/the_saami_culture_archive. [Online; accessed 2-December-2015]
6. Giellatekno, University of Tromsø: North Saami dictionaries (2015). URL http://dicts.uit.no/smedicts.eng.html. [Online; accessed 2-December-2015]
7. Giellatekno, University of Tromsø: Programs for analysing North Saami (2015). URL http://giellatekno.uit.no/cgi/d-sme.eng.html. [Online; accessed 2-December-2015]
8. Grönroos, S.A., Jokinen, K., Hiovain, K., Kurimo, M., Virpioja, S.: Low-resource active learning of North Sámi morphological segmentation. In: Proceedings of First International Workshop on Computational Linguistics for the Uralic Languages. Tromsø (2015)
9. Grünthal, R., Siegl, F.: Uralilaisten kielten pensasmalli ja arvioidut puhujamäärät (The "bush model" of the Uralic languages and the estimated numbers of speakers) (2012). Department of Finno-Ugric Studies, University of Helsinki
10. Jokinen, K.: Open-domain interaction and online content in the Sami language. In: Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik (2014)
11. Jokinen, K., Wilcock, G.: Community-based resource building and data collection. In: Proceedings of 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU-2014), pp. 201–206. St. Petersburg (2014)
12. Jokinen, K., Wilcock, G.: Multimodal open-domain conversations with the Nao robot. In: J. Mariani, S. Rosset, M. Garnier-Rizet, L. Devillers (eds.) Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialogue Systems into Practice, pp. 213–224. Springer (2014)
13. Kulonen, U.M., Seurujärvi-Kari, I., Pulkkinen, R. (eds.): The Saami - A Cultural Encyclopedia. Suomalaisen Kirjallisuuden Seura, Helsinki (2005)
14. Laxström, N., Giner, P., Thottingal, S.: Content translation: Computer assisted translation tool for Wikipedia articles. In: Proceedings of 18th Annual Conference of the European Association for Machine Translation, pp. 194–197 (2015)
15. Laxström, N., Wilcock, G., Jokinen, K.: Internationalisation and localisation of spoken dialogue systems. In: Proceedings of Seventh International Workshop on Spoken Dialogue Systems (IWSDS 2016). Saariselkä (2016)
16. Leinonen, J.: Automatic speech recognition for human-robot interaction using an under-resourced language. Master's thesis, Aalto University, School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo (2015)
17. Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., Navarretta, C.: The NOMCO Multimodal Nordic Resource - Goals and Characteristics. In: Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC 2010), pp. 2968–2973. Valletta, Malta (2010)
18. Palismaa, M., Eira, I.M.G.: Gielas gillii, mielas millii 9 - Davvisámegiela suopmanat (From language to language, from mind to mind 9 - The dialects of North Sami). Davvi Girji, Kárášjohka
19. Sammallahti, P.: The Saami Languages: An Introduction. Davvi Girji, Kárášjohka (1998)

20. Simple4All Consortium: Simple4All: developing automatic speech synthesis technology (2015). URL http://simple4all.org/. [Online; accessed 2-December-2015]
21. Soria, C., Mariani, J., Zoli, C.: Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages. In: Proceedings of XVII FEL Conference. Ottawa (2013)
22. Wikipedia: Motivation (2015). URL http://wikipapers.referata.com/wiki/Motivation. [Online; accessed 2-December-2015]
23. Wilcock, G.: WikiTalk: A spoken Wikipedia-based open-domain knowledge access system. In: Proceedings of the COLING 2012 Workshop on Question Answering for Complex Domains, pp. 57–69. Mumbai (2012)
24. Wilcock, G., Laxström, N., Leinonen, J., Smit, P., Kurimo, M., Jokinen, K.: Towards SamiTalk: a Sami-speaking Robot linked to Sami Wikipedia. In: Proceedings of Seventh International Workshop on Spoken Dialogue Systems (IWSDS 2016). Saariselkä (2016)