COMMUNITY-BASED RESOURCE BUILDING AND DATA COLLECTION

Kristiina Jokinen[1] *and Graham Wilcock*[2]

[1]Institute of Behavioural Sciences
[2]Department of Modern Languages
University of Helsinki, Finland
{firstname.lastname}@helsinki.fi

ABSTRACT

The paper describes our work on participatory and community-based resource collection for the Sami language. This includes community events where participants wrote new Sami Wikipedia articles and took part in speech data collection by reading aloud Sami Wikipedia articles and discussing freely in group conversations. The aim was to increase the number of Sami Wikipedia articles and thereby strengthen Wikipedia as a digital resource for the Sami language and to collect speech data to be used in developing Sami speech components. Such components are intended to be combined with the Sami Wikipedia in order to build a spoken interactive knowledge access system.

Index Terms — language resources development, Wikipedia, Sami language, community-based participatory data collection

1. INTRODUCTION

The new paradigms in communication technology affect language communities in various ways. Social media applications provide novel practices to bring people together, and people's information needs are being met from online collaboratively edited material such as Wikipedia. As language is a means to provide information, its role in the new situations is prominent, and it needs to adapt itself to digital changes. This includes the usual linguistic changes such as lexicon, morphology and syntax, but has an effect also on the level of interaction itself. As [8] notes, language has become "a function that is performed digitally".

This paper discusses aspects of this development, and focuses on the data collection activities for the North Sami language in the DigiSami project. The collection of speech and multimodal video data contribute to building Sami language resources for speech and language technology, while supporting the planning and writing of new Sami Wikipedia articles strengthens the digital text resources for the Sami language. The speech data will be used to support development of speech technology components that can be combined with the Sami Wikipedia to build a spoken interactive knowledge access system.

The paper is structured as follows. Section 2 sets the starting point for the present study by introducing the DigiSami project and the WikiTalk application. Section 3 briefly describes the Sami languages, and Section 4 reviews participatory community-based research methods. Section 5 details the data collection events, participant recruitment, and the three types of data collected: text, speech, and multimodal data. Finally, Section 6 provides discussion and concludes with future directions.

2. UNDER-RESOURCED LANGUAGES AND LANGUAGE TECHNOLOGY

Under-resourced languages typically do not have advanced natural language processing tools such as part-of-speech taggers and syntactic parsers, or sophisticated knowledge resources such as Wordnets and semantic web ontologies. A serious consequence of this is that language technology applications that require such tools and resources cannot offer much support for these languages. It is desirable to find ways to start developing the missing resources for as many languages as possible, and many activities have been initiated for this (e.g. the LRE Map [1] for sharing language resources, or making speech corpora freely available for TTS development [18]). Suggestions for overcoming the many challenges facing small languages include using the same technology standards as big languages, using crowdsourcing and connecting with other small language communities [17].

It is also important to consider what kinds of language technology applications can work successfully with minimal NLP tools or without them. One sophisticated knowledge resource that does not depend on advanced NLP tools for a language is Wikipedia (http://www.wikipedia.org/). Many languages that are under-resourced in language technology tools have their own Wikipedia. The reason for this is that Wikipedia pages can be written and edited by ordinary people with no training in NLP, and the underlying software running Wikipedia is made freely available for all languages by the Wikimedia Foundation (http://www.wikimedia.org/). This means that LT applications that use Wikipedia as a primary resource are potentially able to work successfully even with languages for which sophisticated NLP tools are not available.

2.1. WikiTalk

In earlier research [5, 20] we developed WikiTalk, a speech-based dialogue interface to Wikipedia, through which the user can navigate around Wikipedia following links to any desired topics, and listen to chunks of interesting articles being read out aloud on demand. One version of WikiTalk runs on a humanoid robot [2, 6], so that the robot can act as a talkative conversational companion, able to talk in a very well-informed manner about more or less any topic, while following the human partner's changing interests.

WikiTalk is intrinsically multilingual, as it can work with any language if there is a Wikipedia in the language and if there are speech components for the language. Sophisticated NLP tools such as part-of-speech taggers and syntactic parsers can be used if they are available, but they are not required because the WikiTalk application can use the actual sentences of the Wikipedia articles directly.

2.2. The DigiSami Project

The DigiSami project (http://www.helsinki.fi/digisami/) aims to support digital content generation for the Sami languages with the help of language technology and it focuses especially on the Northern Sami language. It is one of the projects in a joint research framework funded by the Academy of Finland and the Hungarian Academy of Science. The goal of the research framework is to increase the visibility and use of small Finno-Ugric languages in the digital world, as well as to apply language technology tools to develop and process digital material.

At the time of writing, the grand total of articles in all Wikipedias (in all languages) is about 31 million. English Wikipedia is the biggest with 4.5 million articles. The Sami Wikipedia (in Northern Sami) was started in 2004 and has between 7500 and 8000 articles. This places it in the middle of all Wikipedias in terms of size, being ranked 134th by number of articles. For comparison, the Finnish Wikipedia has about 350,000 articles and is ranked 20th.

Sami Wikipedia (http://se.wikipedia.org) contains about 400,000 words in total. The average article length is about 500 characters (as Wikipedias grow, their average article length tends to grow as well). No statistics are available about how many articles were created by translating them from some other Wikipedia.

The extension of applications such as WikiTalk to include Northern Sami will require the necessary resources to be developed. In particular, there is a need for more Wikipedia articles as well as for speech data. To address these needs, the project organized six community events at which new Wikipedia articles were written and spoken language material was collected through video recordings.

3. SAMI LANGUAGE CHARACTERISTICS

Sami is a Finno-Ugric language which is related to Finnish and other Baltic-Finnic languages. It consists of a family of nine Sami languages spoken in the Northern Polar Cap situated in the northern parts of Scandinavia, Finland and the Kola Peninsula in Russia. The languages are spread among four countries: Norway, Sweden, Finland and Russia, and the speakers don't necessarily understand each other. A few of the languages have died over the centuries, e.g. the last speaker of the Akkala Sami spoken in Russia died in 2003.

There are around 30,000 to 40,000 speakers of different Sami languages, most of them in Norway. The languages are: Southern Sami, Uume Sami, Pite Sami, Lule Sami, Northern Sami, Inari Sami, Skolt Sami, Kildin Sami and Ter Sami. Northern Sami is the biggest group with about 20,000 speakers and three main dialects: Torne Sami, Finnmark Sami and Sea Sami. It is also a lingua franca, probably because of the number of its speakers. Our project focuses on Northern Sami, and in this paper, we often use "Sami language" to refer to Northern Sami.

It must be noted that there are more ethnic Sami people than there are speakers of the languages. However, digital information technology has increased interest in the Sami language, for example Wikitravel (http://wikitravel.org) offers Sami-English translations for tourist purposes. One of the important goals of the DigiSami project is to encourage the Sami people to use their language in digital media, but also to increase its visibility among the languages of the world. It is thus important to revitalize the language by encouraging its use in digital contexts, but also to make it visible and available for those who are interested in the language and culture of the Sami people.

Several of the Sami languages have some language technology tools. For instance, Northern Sami has a text analyzer, wordlists, and translation tools (more information at http://giellatekno.uit.no/), and there is a Northern Sami Wikipedia created under the auspices of the Wikimedia Foundation. However, as mentioned above, many of the Sami Wikipedia articles are short compared with the English or Finnish counterparts, and there are not so many links to other articles which are crucial for further navigation and information search among the articles.

4. COMMUNITY-BASED RESOURCE BUILDING

The community effort supported by the DigiSami project to create and write Wikipedia articles is intended to encourage people to develop the Northern Sami Wikipedia more. It is also expected that this kind of community effort will give a boost to the speakers of other Sami languages and other under-resourced language communities to develop their own Wikipedia. Such activity will not only support the Wikipedia encyclopedia, but can be claimed to reap benefit also for language learning and language use. In fact, within the context of multilingual language learning, it has been pointed out [14] how pedagogical methods emphasizing the speaker's own activity and experience provide good results in children's language learning. Concerning minority language learners, it is crucial that there is a possibility to use the language in situations which the speakers find

interesting and relevant to their own activities. This helps them not only to practice with the language, but to act as agents who invent and adapt language practices in the multilingual dynamic environment.

The main idea of Wikipedia is that its knowledge is based on collaborative effort, and the articles can be edited and developed through time by anyone. The structure and information can be discussed by the developers, and the edit history of the pages is visible. It is exactly this kind of community effort that makes Wikipedia what it is: a community built encyclopedia which can improve its contents over time. The recommendations for article writing guide the articles to be generally accepted by the community as topics for the encyclopedia, rather than advertise or argue for a certain viewpoint or blog about oneself. The content of the articles should also be neutral, impartial, and consensus oriented, and it should present the multitude of viewpoints. In particular, the articles should avoid stating that one viewpoint is true or better than another one, and if the topic is controversial, the article should bring forward the opposing and conflicting views, too. The article should also present minority views, and all viewpoints should be presented in a positive and sympathetic way.

From the perspective of Sami language and culture studies, the Wikipedia approach seems attractive. The principle of creating digital knowledge about human cultures by the community members themselves accords with the recent methodological approaches to cultural studies, which emphasize the change from an outside observer viewpoint to active participation in knowledge creation. Such participatory research methods are opposite to those which aim to avoid appraisal, and they deal with the questions of research ethics [7, 9].

It is known that paradigms affect our world views and contain the methodology as well as values. It has often been pointed out (as discussed in [7, 9]) that the main question in indigenous culture studies deals with the different understanding by minority and majority members of what constitutes knowledge. Aboriginal people have their own epistemological and ontological views, and it is emphasized that a scientific paradigm may not only comprise of a theory or a hypothesis that the theory predicts, but enable theory building and hypothesis formation through interaction within the community members. This means that research is not an individualistic process through which the researcher observes and evaluates the community as in the colonial and imperialist traditions, but its starting point is in the community members and in their knowledge which is produced in collaboration with the other members. There is thus a need for new approaches to understand and describe the scientific paradigm, as well as the power relations in the society, although it has also been claimed that a common methodology would be impossible, as the basic ontology is not the same for all aboriginal people [13].

Sami research is also expected to bring benefits to the Sami communities or enlighten others about Sáminess [11]. According to Koskinen [9], the most important aim of indigenous studies is advancing the indigenous identity and self-determination of the indigenous peoples, and she points out that many Sami researchers openly express that the goal is nation-building.

In the DigiSami project, the research goal is to study the effect of digital technology on the Sami language and culture, and the method is hybrid in that it combines language and speech technology with participatory activity. While the data will be analysed using manual annotation and language technology tools, it is expected that this kind of computational account, together with the planned community events, will increase interest in the Sami language, both as a mother tongue and as a second language, and contribute to the revitalisation of the Sami language and characteristics of the Sami culture in the digital era.

5. DATA COLLECTION

The data collection and Wikipedia article development were organized through community events where participants were invited to take part in three different tasks. First, the participants discussed possible Wikipedia topics and wrote the selected contents down as articles. Second, the events consisted of audio recording of Wikipedia text reading, and finally, there was video recording of free conversations. The tasks contributed to the increase of Wikipedia articles and to speech data collection on read text and conversational speech. Below we briefly describe the setup, participants, and the collected corpora, and refer the reader to [4] for further details as well as for a preliminary analysis of the corpus.

5.1. Setup

The community events took place in four villages in Finland: Enontekiö, Utsjoki, Inari and Ivalo, as well as two towns in Norway: Kautokeino and Kaarasjoki. The events were organized at high schools and in community halls and libraries. The school events were for pupils who participated in the writing as part of their mother tongue education, while the other events were meant for those working during the day. The locations were selected so that the main Sami speaking areas were represented with different Northern Sami dialects, and the main Sami central towns were also represented. The collection concerns only Northern Sami.

The participants were asked to take part in three different tasks: Wikipedia article planning and writing in the Sami language, Wikipedia article reading aloud in the Sami language, and conversation in the Sami language. Data collections were organized in parallel to Wikipedia writing sessions. Due to personal information being recorded in video and speech data, the participants (or their parents if under-aged) were asked to sign a data usage agreement where they explicitly allow the data collection and its use for research purposes in the project.

The length of the whole event was about three hours, and the general structure of the events was the same. First the

participants were introduced to the project and the tasks. The Wikipedia planning and writing took place in groups of 2-3 members, and the video and speech recordings were parallel to the text writing. The groups came to video recordings together, whereas the text readings were done one person at the time. If the participants had questions or comments at any point during the event, they could ask the instructors who were ready to help with the writing and other issues. Also, if the participant wanted to finish, they could do so by contacting the instructor. All participants finished their article writing, but with text reading some participants opted to read only one of the three requested articles, while some were willing to read extra texts as well.

5.2. Recruiting of participants

Many different means were used to recruit participants for the events. High schools, colleges, universities, libraries, museums, etc. were contacted by phone and e-mails. Also the Sami parliament was contacted, and an advertisement of the events was published in the local newspaper *Enontekiön Sanomat*. The Sami language regional TV-news YLE Sápmi interviewed the project leader and one of the participating teachers about the project's goals, thus giving a positive push for the forthcoming collection events. Exploiting the new social media, the project also used Facebook to contact various Sami Facebook groups as well as private people. The project was promoted by joining in different Facebook groups of Sami speaking people, and posting on their wall.

A social media effect was clearly seen in the recruitment in that the contacted people suggested their friends and colleagues and other people whom it would be worthwhile to contact. Many contact people appeared being in contact via different sources and they received information about the project from many different sources. Altogether we contacted about 50 people who then spread the word.

An important role was played by the Northern Sami teachers at the upper secondary schools in Utsjoki, Ivalo and Karasjok: it was possible to agree to have the events at the school during mother tongue lessons, and organize the data collection journey around them.

5.3. Text corpora

The goal of the Wikipedia writing sessions was to collect as many Wikipedia articles as possible so as to increase the number and quality of the current Wikipedia articles. Each group could select a topic or topics they wanted to write about, but they could also choose to extend an existing article or translate another Wikipedia article into the Sami language. The articles were to be written according to the Wikipedia recommendations (see above), and they were expected to be fairly complete content-wise, written in North Sami and following the written language conventions. We anticipated that the final editing and formatting could be done by the instructors, but in fact, the groups finished with complete Wikipedia articles. It is good to notice, however, that Wikipedia articles can be later edited so as to update and modify the information as is necessary. For instance, during the short time for the community events, there was not enough time to find the important references, but these can be added later on.

The participants could freely discuss with other groups about their Wikipedia texts and get feedback about them, and also give feedback and comments on the other's texts. The topics of the Wikipedia article related to the Sami language and culture: traditional food, clothes, houses, games and pastime, social life and livestock, reindeer herding, fishing, etc., even one about Tolkien.

The Sami Wikipedia will function as the text corpora for further language technology applications, and in particular, will form the basis for a future Sami version of WikiTalk.

5.4. Speech corpora

The read speech audio corpus consists of each participant reading aloud three Sami Wikipedia articles. The reading took place in a calm and normal speaking manner, and the participant could study the articles in advance. The readings were recorded but not videotaped.

The speech samples are from 28 participants, of which 10 are men and 18 women. Their age range from 16 to 65 years: 17 were 16-21 years old, five 30-44 years old, and six 49-65 years old. All but one of the participants were native speakers of Northern Sami. One male participant had learnt Northern Sami as a second language, but he was a fluent speaker and used Sami daily at work.

5.5. Multimodal corpora

The multimodal corpus includes recordings of eight natural conversations. Each group took part in one conversation of about 10-15 minutes long (provided that the participants had given their consent to be videotaped). The group sat around a table so that they could see each other well, and their task was to discuss Wikipedia article writing and issues related to the Sami language, although they were also encouraged to discuss whatever topics that would interest them, so as to support free and natural interaction.

The conversations were videotaped by two Panasonic HC-X920 video cameras and three GoPro HERO3 cameras so that each camera pointed at each participant and one recorded the whole situation. The reading was recorded by EDIROL R4Pro four channel recording device with AKG 417 L microphones. The conversations were also recorded by the same device, besides the camera's own microphone.

In Kautokeino and Inari where the events were organised in a public community hall, the number of participants was limited, and it was difficult to organise a conversation, so it was decided to collect read speech only.

6. DISCUSSION AND FUTURE WORK

The response rate among ethnic minorities has long been a problem within socio-cultural studies. Feskens et al. [3] suggest that to increase the response rates, the researchers

should focus on enhancing the contact rate and reducing the number of non-respondents who are unable to produce the required information. We also found it difficult to get people interested in participating in the events, even though the events were meant to be community events which were beneficial for the whole community. Although our contact persons were very cooperative and suggested several other people for us to get in touch with, many of them were not so eager to participate in the events themselves.

The first contacts were made about three months in advance, and some contacts continued through till the actual data collection trip. Despite the several channels of advertising and contacting people (email, telephone, TV, newspaper, Facebook), especially close to the time of the events, the turnout was lower than expected, and it was not clear if there should have been more personal contacts on the very day of the event so as to draw people in. In the community halls, it was the personal effort by the instructors and the staff that brought participants in.

While there may be several good reasons for low turn-up and interest rates, starting from the busy time or unsuitable event schedule, we discuss two a little more: the person's own internal constraints, and the culturally-conditioned general situation.

It is possible that people may have found the situation to write Wikipedia articles difficult and intimidating. Although we did not expect Wikipedia articles to be completed within the three hour session, and explicitly offered to do the final editing and formatting ourselves, the idea itself of writing Northern Sami texts may have appeared haunting: maybe the people did not feel comfortable in writing an article, nor confident in spelling their language. Moreover, to do this for some fancy language technology goal might have sounded odd or problematic. The options to start with an existing article and expand its content, or just to discuss about possible content of an article and let us to write the thoughts down, may not have come through as suitable alternatives.

Another reason may be related to the history of the Sami language which was suppressed heavily until very recently. The Sami culture has, indeed, been studied for more than 300 years [10], but the Sami people have gained their rights, and the self-value and use of their language only during the last couple of decades. The issue may thus concern the ownership of the Sami language and the Sami culture, and be related to the question of research methodology discussed in Section 4. Community-based research paradigm, even though inviting to conduct studies on one's own language in collaboration and interaction with the other community members, may still require initiation from inside: this way the research could be free from the old imperialistic "avoid appraisal" approach that has long prevailed in cultural studies, and become part of the active creation of the culture and language at the situation itself.

The last point to consider in community-based resource collection is the focus on Wikipedia itself. One of the most important guidelines in creating Wikipedia articles is the requirement of the neutral status of the information in the articles: the articles should avoid first-hand experiences and opinions which cannot be referenced. However, when writing about the indigenous culture and heritage, problems may be encountered concerning subjectivity and the value given to personal experience and shared understanding that emerge from community-based activities. What makes this a problem for studies is that, although there is also plenty of written history dating back to the seventeenth century, a lot of the indigenous Sami culture has been passed on orally from one generation to the next. To capture these aspects, other techniques, such as story-telling, can be useful. We are looking for developing digital solutions, such as described in [15], to preserve this kind of intangible cultural content in sharing of experiences. For instance, we can collect recording of audio narration, videos and interviews that preserve the traditional practices regarding Sami lifestyle and personal experiences of marginalization, and they can be presented using the WikiTalk presentation techniques.

ACKNOWLEDGEMENTS

We thank Hanna Kellokoski and Jani Koskinen for their work in organizing the community events and processing the collected speech data, and Niklas Laxström for providing the statistics about Sami Wikipedia. We also thank the teachers, students and all the participants in the Wikipedia article writing and data collection events.

REFERENCES

[1] Calzolari N., Del Gratta R., Fracopoulo G., Mariani J., Rubino F., Russo I. and Soria C. The LRE Map. Harmonising Community Description of Resources. *Proceedings of LREC 2012,* Istanbul, Turkey, pp. 1084—1089, 2012.

[2] Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K. and Wilcock, G., Multimodal conversational interaction with a humanoid robot. *Proceedings of CogInfoCom 2012*, p. 667-672, 2012.

[3] Feskens, R., Hox, J., Lensvelt-Mulders, G., and Schmeets, H. Collecting Data among Ethnic Minorities in an International Perspective. *Field Methods*, Vol. 18, No. 3, p. 284–304, 2006. DOI: 10.1177/1525822X06288756

[4] Jokinen, K. Open-domain Interaction and Online Content in the Sami Language. *Proceedings of the Language Resources and Evaluation Conference (LREC 2014).* Reykjavik, Iceland, 2014.

[5] Jokinen, K. and Wilcock, G. Constructive Interaction for Talking about Interesting Topics. *Proceedings of the Language Resources and Evaluation Conference (LREC-2012)*. Istanbul, Turkey, 2012.

[6] Jokinen, K. and Wilcock, G. Multimodal Open-domain Conversations with the Nao Robot. *Proceedings of the Fourth International Workshop on Spoken Dialogue Systems (IWSDS 2012),* Ermenonville, France, 2012.

[7] Keskitalo, P, Määttä, K. and Uusiautti, S. Ethical Perspectives on Sámi School Research. *International Journal of Education*, Vol. 4, No. 4. pp. 267-283. 2012.

[8] Kornai, A. Language Death in the Digital Age. Invited lecture at META-FORUM 2012 - A Strategy for Multilingual Europe, 20 June 2012, Brussels. Available as a video lecture at http://videolectures.net/metaforum2012_kornai_language, 2012.

[9] Koskinen, I. Critical Subjects: Participatory Research needs to Make Room for Debate. PSA 2012 Biennial Meeting, 2012.

[10] Kulonen, U.-M., Seurujärvi-Kari, I. and Pulkkinen, R. (eds.). *The Sámi: A cultural encyclopaedia.* Helsinki, Finland: Suomalaisen Kirjallisuuden Seura. 2005.

[11] Länsman, A.-S. Kenelle saamentutkija tutkii? [For whom does a Sámi researcher do research?] In K. Lempiäinen, O. Löytty, & M. Kinnunen (Eds.), *Tutkijan kirja [A researcher's book]* pp. 87–98. Tampere: Vastapaino. 2008.

[12] Mörth, K., Declerck, T., Lendvai, P. and Váradi, T. Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts. In: E. Montiel-Ponsoda, J. McCrae, P. Buitelaar, P. Cimiano (Eds.) *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web,* Bonn, Germany, Springer, 2011.

[13] Oskal, N. The question of methodology in indigenous research. A philosophical exposition. In H. Minde, S. Jentoft, H. Gaski & G. Midré (eds.) *Indigenous peoples: self-determination, knowledge, indigeneity*. Eburon: Delft, pp. 331–345. 2008.

[14] Pitkänen-Huhta, A. and Pietikäinen, S. Toiminnallisuus kielenoppimisessa - pedagogisia kokeiluja saamen luokissa (Agency in Language Learning - Pedagogical Experiments in Sámi Language Classes). *Kieli, koulutus ja yhteiskunta* 10/2013. Available at: http://www.kieliverkosto.fi/article/toiminnallisuus-kielenoppimisessa-pedagogisia-kokeiluja-saamen-luokissa/

[15] Rodil, K. A Participatory Perspective on Cross-Cultural Design. In Ebert, A., van der Veer G.C., Domik, G., Gershon, N.D., Scheler, I. (eds.). *Building Bridges: HCI, Visualization, and Non-formal Modeling*, Springer Lecture Notes in Computer Science, pp. 30-46. 2014.

[16] Seurujärvi-Kari, I., Pedersen, S. and Hirvonen, V. The Sámi. The indigenous people of northernmost Europe. *European Languages 5*. Brussels: European Bureau for Lesser Used Languages. 1997.

[17] Soria, C., Mariani, J. and Zoli, C. Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages Proceeding of XVII FEL Conference 10/2013, Ottawa

[18] Stan, A., Watts, O., Mamiya, Y., Giurgiu, M., Clark, R., Yamagishi, J., and King, S. TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision. Proceedings of Interspeech 2013, 2013.

[19] Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. and Nagy, V. Parallel corpora for medium density languages. In: *Recent Advances in Natural Language Processing (RANLP 2005),* pp. 590–596. 2005.

[20] Wilcock, G., WikiTalk: A Spoken Wikipedia-based Open-Domain Knowledge Access System. *Proceedings of the COLING 2012 Workshop on Question Answering for Complex Domains*, Mumbai, India, pp 57-69, 2012.