Open-domain Interaction and Online Content in the Sami Language

Kristiina Jokinen
University of Helsinki
Institute of Behavioural Sciences
Siltavuorenpenger 1 A, 00014 Helsinki
E-mail: kristiina.jokinen@helsinki.fi

Abstract

This paper presents data collection and collaborative community events organised within the *DigiSami* project concerning the North Sami language. The *DigiSami* project is one of the collaboration initiatives on endangered Finno-Ugric languages, supported by the larger framework between the Academy of Finland and the Hungarian Academy of Sciences. The goal of the project is to improve digital visibility and viability of the targeted Finno-Ugric languages, as well as to develop language technology tools and resources in order to assist automatic language processing and experimenting with multilingual interactive applications.

Keywords: spoken language corpora, Wikipedia, interaction engagement, North Sami language, community collaboration

1. Introduction

The project *Finno-Ugric Digital Natives: Linguistic support for Finno-Ugric digital communities in generating online content* is a four-year collaboration project between University of Helsinki and the Research Institute of Linguistics in Budapest, and its goal is to support digital visibility and viability of the minor Finno-Ugric languages. The project is one of the joint initiatives within a larger framework between the Academy of Finland and the Hungarian Academy of Sciences, on endangered Finno-Ugric languages, and it aims to address needs of the small Finno-Ugric language communities in the current globalizing world. The partners have their own research agenda but they can share tools, resources and expertise available at their research sites.

Digital viability of the target Finno-Ugric languages is addressed in two ways: (1) by generating online content on the target languages, and (2) by developing language technology tools and resources for automatic language processing and further development of interactive applications. The work in Helsinki explores especially WikiTalk, an open-domain dialogue application that allows the user to navigate among Wikipedia articles (Csabo et al., 2012; Jokinen and Wilcock, 2012), and how the system can be adapted and used for integrating social media discourse features in multilingual human-robot interactions. Linguistic support will be carried out in the framework of standardized, linked infrastructure, to allow for interoperability, machine reading, and sustainability.

This paper presents data collection and collaborative community events organised by the Finnish partner within the above framework. The fieldwork concerns the North Sami language, and thus the project is abbreviated as *DigiSami* (DigiSámegiela). Data collection consists of collecting a North Sami spoken language corpus including both read and conversational speech. Another crucial goal of the collection exercise is to encourage participants to write articles for the North Sami Wikipedia (either new ones, or extend stubs). This activity is not restricted to a particular event, but can, of course, be done by individuals themselves later on. The importance of the particular events organised by the DigiSami project is simply to raise awareness among the speakers of such an activity for the purpose of strengthening digital visibility of the Sami language, and consequently revitalise its use in internet.

North Sami (Davvisámegiela) is one of the nine Sami languages spoken in the northern parts of Scandinavia, Finland, and the Kola Peninsula in Russia. There are about 40 000 speakers of the different Sami languages, and North Sami is the biggest group of about 20 000 speakers, and also functions as a lingua franca. We have chosen North Sami as our target language for the simple reason that there already exists a North Sami Wikipedia, and because several language technology tools have been developed which can be used for online content analysis and development of interactive applications within the general goals of the project. It is possible to apply the project work to other Sami languages too, given that the necessary infrastructure and tools are available (e.g. in case of WikiTalk, the existence of Wikipedia in that particular language is needed). There are already several tools developed and openly available for the various Sami languages (and other small Finno-Ugric languages) on the website of Sámi Giellatekno (Centre for Sami language technology, Arctic University of Norway, Tromsø), see: http://giellatekno.uit.no/english.html

The rest of the paper is structured as follows. Section 2 will motivate the research and discuss the issues encountered by endangered languages in the digital era. In Section 3, relevant features concerning interaction with the WikiTalk open domain dialogue system are briefly presented. Data collection for the North Sami language and the results are described in Section 4. Section 5 then provide conclusions and discussion for future work.

2. Digital Era and Globalized language use

Internet and digital information have brought in new paradigms for communication. Not only it is possible to maintain fast and efficient immediate communication, but it is also possible to immerse in social interactions with partners who are not physically co-present, via the new type of media applications such as Facebook, Twitter, YouTube, Wikipedia, and other Wikimedia-related

initiatives. Novel forms of communication have emerged to offer virtual co-presence anytime and anywhere. Social media are becoming extremely popular among any speech community (which has internet connection), and local versions of social media appear regularly, indicating the need for localized versions of media applications.

All this have had, and continues to have, profound impact on language communities. On one hand, people who speak different languages are brought together through various social media platforms to get connected and work on their joint interests, and on the other hand, internet has gradually established itself as an open knowledge source, so that the speakers' information need is covered online. Concerning the latter, pertinent issues deal with the origin and reliability of the information as well as with the issues related to openness and availability of the information for everyone (who has internet access), privacy issues, and ethical considerations. An important characteristic of user-generated online materials is also the sense of community: material is collectively produced and often edited by the users, so each individual has an important role to play as an author and an editor.

Collaboratively edited articles have also prompted critical discussions about what constitutes "objective" information, and consequently emphasised the relative nature of the "truth". Open material is based on interpretations, assumptions, values, and attitudes of the authors in a dynamically changing environment, and its presentation is always done from a certain perspective representing a certain set of values. Even though it may aim to be as general and objective as possible, it still remains relative to the author's view of the true state of affairs, while some user-generated content may be related to a group of authors who entertain similar views, and include opinions which are not globally acceptable.

Established knowledge sources such as Wikipedia, have no rules as such about how to write articles, but encode best practises in their recommendations (see e.g. https://fi.wikipedia.org/wiki/Wikipedia:Viisi_pilaria). For instance, the recommendations emphasise in the very beginning that the articles should be written from a neutral perspective which pays equal attention to various controversial aspects and view-points; this is to guarantee the encyclopaedic nature of Wikipedia as a knowledge source that provides accepted information. For this purpose it is important that the articles contain established information rather than scientific hypotheses, and they should also have references to reliable sources whenever possible. In particular, Wikipedia article should not be a collection of trivial points or a discussion forum for personal opinions, experiences and statements. An important aspect of Wikipedia is that it is free material which can be edited, modified and used by anyone, and the collective editing and responsibility of the content of the articles is a good guarantee for the acceptability and factual truthfulness of Wikipedia content.

The new modes of interaction also affect speech communities by shifting native language use towards a "globalized" language use: native languages are often "traded" for the majority language, to enable faster and wider communication. The impact of this is likely to be bigger on small language communities than on major language communities, although more research is needed to determine the extent and dimensions of the changes in these speech communities.

Language is used as a vehicle to express and transfer information, emotion, and rapport, and it can obviously adapt itself to inevitable changes occurring in the society and communication. However, as Kornai (2012) points out, digitalization does not only result in adaptation on the lexical level by coining new terms for new concepts, but has more far-reaching tendencies in that the language has "a function that is performed digitally". In other words, it is not enough that a language has a passive presence in the web (i.e. it is mainly used to read texts, for which it is enough to maintain lexicons, literature, news services, etc.), but the language also has to produce new, publicly available digital material. A language can survive in the current global economic and information space 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.

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. In particular, the project

- inspects, analyzes, and characterizes the language use in social media. For this purpose, data from dialogue-related genres will be collected and annotated on levels ranging from grammatical up to discourse phenomena.
- experiments with language technology applications that can strengthen the user's interest in using the language in various interactive contexts. For this purpose, focus will be on the WikiTalk system and the extension of its capabilities to study feasible interaction strategies in multilingual context.
- aims to alleviate barriers in accessing information from user-generated content. For this purpose, community-based generation of translated material on the web will be supported, based on partially existing language resources and technology concerning comparable corpora, cf. Váradi and Héja, (2011); Varga et al. (2005).

3. Nao WikiTalk

The Nao WikiTalk is an automatic dialogue system with which users can interact and talk about interesting topics, as well as navigate through the Wikipedia articles looking for information (Jokinen and Wilcock, 2012, 2013). The system uses Wikipedia as a knowledge base for open-domain dialogues, and it can thus be adapted to any language which has their own Wikipedia. The application allows experimentation with various discourse level issues related to information flow and coherence of presentation. Using a situated agent, it is also possible to study multimodal communication issues ranging from gestures to gaze and body posture, in order to enable natural and

expressive presentation of the content. For instance, the version of WikiTalk implemented on the Nao-robot (Csabo et al., 2012) focusses on the robot's gesturing and how gestures can be used to signal new information (beat gestures). The gestures clarify spoken presentation by indicating what the links of the article are, which the user can choose to follow. The Nao WikiTalk can be adapted to different scenarios where the robot has the role of a companion and provides the user with interesting and useful information from online services. It can also be used in educational settings where the interactive system is used as a learning tool (Jokinen & Majaranta, 2013).

In the context of DigiSami, the plan is also to collect speech resources for North Sami, so as to enable building of WikiTalk -type interactive applications. Also existing language technologies are taken into account when modelling novel forms of language used for displaying and conveying information on social forums (see Lendvai, (2011), Mörth (2011)). For instance, by annotating lexical chains and topic maps in comparable corpora, it is possible to retrieve similar discussions in other languages, and compile parallelized discussion threads. Conceptual relations can also be extracted between user contributions and the WikiTalk system responses, and these can be integrated in the topic knowledge. In WikiTalk, articles and hyperlinks are organised with the help of Topic trees into the context of interaction (cf. Jokinen and Wilcock 2012). Topic trees can be generalised in regard to world knowledge, e.g. WordNet, to provide the users with more freedom and flexibility to express their requests.

4. Corpus Collection

4.1 General Organisation

The project organized special data collection events in the main Sami speaking areas, with the goal of collecting two types of speech data, read speech and conversational speech, as well as encouraging the participants to write Wikipedia articles. The events took place in four towns in Finland: Enontekiö, Utsjoki, Inari and Ivalo, as well as in two towns in Norway: Kautokeino and Karasjok. They were selected to represent the central Sami-speaking areas and the different North Sami dialects. The collection concerned only North Sami, but it was hoped that the events would act as encouragement for similar work on other Sami languages as well.

The six events were organized at high schools and in community halls and libraries. The school events were for students who participated in the data collection as part of their mother tongue education, whereas the other events were meant for other interested participants.

The participants were recruited by approaching people and institutions that were one way or another involved with the Sami language and/or promoting interest and knowledge of the Sami culture and language widely. High schools, colleges, libraries, museums etc. were contacted by phone and emails, as well as people known to be interested in the Sami language or speaking it as their mother tongue. Also the Sami parliament was contacted. The school events were organised by contacting the Sami language teachers, and asking if the event could be organised at the school. The teachers were extremely helpful in the organisation of the events and also during the actual event, supporting the students' tasks and also participating in the Wikipedia article writing themselves.

The events were also advertised in the local newspaper *Enontekiön Sanomat*, and the Sami language regional television YLE Sápmi made a short news item concerning the project's goals and the forthcoming collection events. An interview of the project leader and one of the participating teachers was broadcasted just before the data collection events started.

The new social media, Facebook, was also used to recruit participants: the Sami Facebook group was contacted, and postings placed on the walls of different other groups of Sami speaking people in Facebook. The contacted people were usually very helpful, and although straightforward commitment was rare, they pointed out colleagues and friends who might be interested in the participation, and also offered to advertise the project themselves. Social media effect was clearly seen in the networking efforts: many contacted people appeared to be in contact with several networks and received information about the project from many different sources. Altogether we contacted about 50 people who then spread the word.

4.1 Setup of the Sessions

The general structure of the events was the same although there was variation depending on the location and the number of participants. For instance, at the schools in Ivalo, Utsjoki and Karasjok, the participants were asked to take part in all three Sami language tasks: planning and writing of Wikipedia articles, reading aloud three Wikipedia articles, and taking part in a video conversation with their classmates. 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.

The length of the whole event in one particular place was three hours. In the very beginning of the event, the participants were briefly introduced to the project goals and the tasks. The Wikipedia planning and writing took then place in groups of 2-3 members, each of which selected a topic or topics they wanted to write about. The group also had to agree on how they were going to organize the writing: who was the secretary and wrote the article for the group, or if each would write an article themselves. Besides producing a brand new article, the group could also choose to extend an existing Wikipedia article or to translate another Wikipedia article into the Sami language. The group could freely discuss and plan the content of their article, and there was also internet connection available so the participants could search for information as necessary. The resulting articles were expected to be fairly complete content-wise, written in North Sami and following its orthographic and grammar conventions. The project assistants were available to

provide help in the editing and formatting of the articles, but it turned out that the participants were already rather knowledgeable of the technical aspects of Wikipedia articles.

Data recordings were organized in parallel with the Wikipedia writing session. A participant group of three members (or a participant pair, if the students worked pair-wise) came to the "recording studio" one at the time, and was instructed to converse as naturally and freely as possible on the topic they had been writing, or if they wished, on any other interesting topic related to their hobbies, favorite music, plans for the evening, etc.

After video conversations with all the groups, the recording continued with read speech. The participants were asked to come to the recording studio one by one to read three short texts, while the other group continued their Wikipedia article writing. The texts were three existing Wikipedia articles dealing with the Sami language, the snowmobile, and the Sami traditional dress. The participants were instructed to read in a calm clear and normal manner, and have a pause in between the texts if they wanted. The participants could also read the texts before they read them aloud, so as to become familiar with the content and help the reading to become smooth and fairly fluent. Some participants chose to read only one article, being too tired to read all three. Others, on the other hand, were willing to read more text, and they read the instruction sheet as well. The readings were not video-taped.

4.3 Collected data

The aim of the data collection was to collect data that would be of good quality so that it would be possible to use it for language technology processing, especially for developing speech recognition and speech synthesis applications. In this respect, the collected data is one of the first, if not the first systematic resource for spoken North Sami.

The reading was recorded by EDIROL R4Pro four-channel recording device with AKG 417 L microphones. Also the conversations were recorded by the same device, besides the camera's own microphone. Two Panasonic HC-X920 video cameras and three GoPro HERO3 cameras were used for video-recordings, so as to get a view of each participant and of the whole situation. Figure 1 shows a conversational setup in a paint brush format.

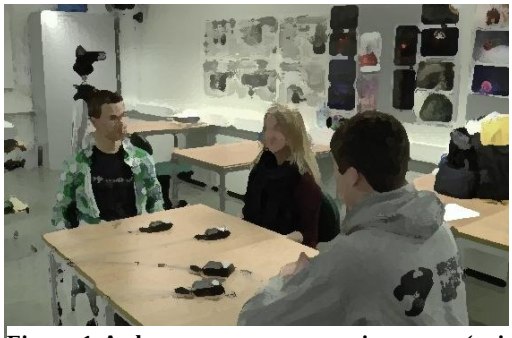

Figure 1 A three person conversation setup (paint bursh format).

All the participants signed a data usage agreement, where they explicitly allow the data collection and the use of collected data for research purposes in the project (those who were under-aged, a signature by one of the parents was requested). The participants also filled in a short ethnographic questionnaire dealing with their age, place of living, migration, use of the Sami language, motivation to take part in the event, etc. The participants were not paid for the participation, but during the event they were offered refreshments and at the end of the event they were given small gifts as tokens of gratitude of their participation.

Altogether we collected eight conversations, each about 10-15 minutes long, as well as read speech from 28 participants. There were 10 men and 18 women who took part in the events, and their age ranged 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 North Sami. One male participant had learnt North Sami as a second language, but he was a fluent speaker and used North Sami daily at work.

All participants were bilingual, and spoke fluently also the majority language, i.e. Finnish, or, in Kautokeino and Karasjok, Norwegian. Most participants had lived their life in the Sápmi area, although not in the same town or village. 10 participants had also lived in bigger cities in the southern part of the area, Oulu and Rovaniemi in Finland and Bergen in Norway, for a short period of time.

26 participants reported they use North Sami daily, while 1 participant reported using North Sami weekly and 1 participant monthly (both of these used North Sami with the family members on the father's side). 89% of the participants use North Sami at home and 75% at school/work, while only 46% use North Sami in shops, offices, and restaurants. It is interesting that the participants seem to use North Sami in their main daily activities when communicating with family members, teachers and co-workers, but less than half used North Sami when communicating in other social situations. This may reflect the fact that the interlocutors in shops and offices are often majority language speakers. Another interesting observation is that 57% (16/28) of the participants said they use North Sami with all the people they communicated with, i.e. both family members and outsiders, while 2 participants reported they use North Sami when communicating with teachers and other Sami people, i.e. with outsiders, not with family members, and the non-native speaker used North Sami at work only. 32% (9/28) of the participants use North Sami with family members from either side only: 4 on the mother's and 5 on the father's side.

Concerning the participants' schooling in North Sami, 21 participants (75 %) have had their basic education in North Sami, half of the participants have had their upper secondary school education in North Sami, and 11 had taken the North Sami examination within the national matriculation examination (A-levels) at the end of the upper secondary school.

The conversations were fairly natural, although significant puzzlement and shyness was also observed and affected e.g. the number of utterances and turns, topic introduction, and general spontaneity. However, the main purpose of the conversational data collection was to get speech data that is produced in conversational setting rather than to provide data for conversation studies, so the interactional behavior was not considered a problem as such.

The recorded data is currently being transcribed and anonymised, and the Wikipedia articles are already on the web. The topics of the produced Wikipedia articles are related to the Sami culture, but one also concerns Tolkien's Lord of the Rings.

5. Conclusions and Future Work

In many new scenarios of communication technology, the user's information need is channelled through online, collaboratively edited material such as Wikipedia, and the user's interaction with such material is conducted via social media platforms and interactive automatic applications. Language communities are sensitive to these new paradigms, but without appropriate tools, such communication in one's own language may become impossible. In fact, in the digital age, a language may become endangered by not having appropriate language technologies at its disposal.

The DigiSami project, which is conducted in the larger framework concerning minor Fenno-Ugric languages in the digital world, addresses these challenges with respect to North Sami. The two general goals of the project are: (1) to collect North Sami speech data and provide necessary material for building interactive applications with the help of available language-technology methods and tools, and (2) to advocate the use of North Sami in collaborative editing and content creation, especially in digital contexts such as Wikipedia. It must be emphasised that in this way the project will not only aim to generate new knowledge of the Sami language and culture, but simultaneously to contribute to the North Sami language revitalization, by supporting its active use in interactions with digital contexts.

The project organised a data collection trip to the Sápmi area for these purposes. The trip was encouraging and successful in many ways. First, the collected North Sami speech data is a valuable resource as a systematic and comparable spoken North Sami corpus. Although small, it provides a starting point for further studies on the North Sami spoken language characteristics and speech technology. Second, another positive result is that the trip promoted collaborative community work and in particular, it encouraged the participants to improve the North Sami Wikipedia by writing more articles in it. Moreover it was hoped that the organised community events would encourage speakers of the other Sami languages also to establish Wikipedia in their own language, and indeed, some Inari Sami (*anarâškielâ*) speakers were interested in learning more about how to establish and maintain Wikipedia in their language, thus initiating truly community-based approach to Wikipedia content production. This would allow more Sami people to take part in the development of Wikipedia and contribute to the digital visibility of their language in their own terms, and enable their cultural identity be preserved on the basis of their own experience, rather than be defined from point of view of outsiders. For instance, according to Koskinen (2012), the most important aim of indigenous studies is advancing the indigenous identity and self-determination of the indigenous peoples.

The participants' own comments on the event were also supportive, and pointed to the same direction as the general project goals. While most of the participated students took part in the events due to their obligatory school curriculum, many also expressed their interest in the project, as well as in the Sami language and the WikiTalk application. Two participants wrote that they were curious about the project and the new technical applications, while one was wondering what benefits the project may bring to the Sami language. Four participants wrote explicitly that it is good to be involved in all the activities that help the Sami language, and in this case, their motivation to take part in the event was based on the promotion of the Sami language in the digital world. Wikipedia was also regarded as an important tool, and in fact, it turned out that at least one of the participants had already earlier written some Wikipedia articles.

Finally, the data collection team gained deeper understanding of the practical details related to technical requirements of video recording, and in particular, to the Sami language and culture as well as to the planning and organisation of community-based events. These issues are discussed more in Jokinen (2014).

Future work deals with the data preparation for the analysis. This includes transcription and phonetic analysis of the read and conversational speech using Praat (Boersma, and Weenink, 2009), and also some general level interaction analysis of the conversations. This is not planned to be as detailed as presented in Allwood et al. (2007), Paggio et al. (2010), or Jokinen and Tenjes (2012), but include e.g. general topic structure of the dialogues, engagement, and eye-gaze (cf. Jokinen, 2011).

The project framework aims to support Finno-Ugric language communities in the digital world so that they are able to cope with some of the digitally performed functions of their native languages. Data collection and Wikipedia article writing events organised by the DigiSami project are steps towards this direction. Digital visibility is important in enabling and encouraging people to participate in social discussions using their own language, and language technology, targeting various forms of written text and spoken discourse, can become an enabler technology in this context, for applications that help people to collaborate and share knowledge.

6. Acknowledgements

The work is done in collaboration with Tamas Váradi in the Research Institute for Linguistics at the Hungarian Academy of Sciences. Thanks go also to Piroska Lendevai

for her help in the early stages of the project, and to the Finnish team, Graham Wilcock and Niklas Laxström for their work on the WikiTalk application, and Hanna Kellokoski and Jani Koskinen for assisting in the data collection. We are extremely grateful to all the people who spread the word about the project, and helped us in many ways to contact potential participants as well as organize the events. Moreover, we would like to thank the participating schools and their principals for accepting our data collection visits, and in particular, the teachers for helping us before and during the actual event: Marjaana Aikio (Ivalo), Helmi Länsman (Utsjoki), and Mette Anti Gaup (Karasjok). The biggest thanks go to the participants themselves: they produced Wikipedia articles, read aloud three articles, and took part in the video conversations, and without whom the DigiSame corpus would not exist.

7. References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P. (2007) The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In Martin et al. (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*.

Boersma, P. and D. Weenink (2009). Praat: doing phonetics by computer (version 5.1.05). Retrieved May 1, 2009, from http://www.praat.org/

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

Jokinen, K. (2011) Multimodal Information - Collection and Analysis of Interactive Data. HCII 2011. Orlando, U.S.

Jokinen, K. (2014) Community-Based Resource Building and Data Collection. The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14). St.Petersburg, Russia.

Jokinen, K. and Majaranta, P. (2013). *Eye-Gaze and Facial Expressions as Feedback Signals in Educational Interactions*. In D. Griol Barres, Z. Callejas Carrión, R. López-Cózar Delgado (Eds.) Technologies for Inclusive Education: Beyond Traditional Integration Approaches. Chapter 3, pp.38-58. Hershey, PA: Information Science Reference, IGI Global.

Jokinen, K. and Tenjes, S. (2012). Investigating Engagement Intercultural and technological aspects of the collection, analysis, and use of Estonian Multiparty Conversational Video Data. *Proceedings of the Language Resources and Evaluation Conference (LREC-2012).* Istanbul, Turkey.

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

Jokinen, K. and Wilcock, G. (2013). Multimodal Open-domain Conversations with the Nao Robot. In: Mariani, J., Devillers, L., Garnier-Rizet, M. and Rosset, S. (eds.) *Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice*. Springer Science+Business Media

Kornai, A. (2012) 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/

Koskinen, I. (2012) Critical Subjects: Participatory Research needs to Make Room for Debate. *PSA 2012 Biennial Meeting*.

Lendvai, P. (2011). Towards a Discourse-driven Taxonomic Inference Model. In: Bouma, G. and van den Bosch, A. (eds.) Interactive Multi-modal Question Answering, Pages 255-274. Springer, 2011

Mörth, K., Declerck, T., Lendvai, P., Váradi, T. (2011) 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

Paggio, P., J. Allwood, E. Ahlsén, K. Jokinen, C. Navarretta (2010). The NOMCO multimodal Nordic resource - goals and characteristics. In *Proceedings of LREC 2010*, 2968-2973.

Prószéky, G. (2011) Endangered Uralic Languages and Language Technologies. Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 1–2, Hissar, Bulgaria, 16 September 2011.

Suihkonen, P. (1998) Documentation of the Computer Corpora of the Uralic Languages at the University of Helsinki. Technical Reports TR-2. Helsinki: Department of General Linguistics, University of Helsinki.

Váradi, T., Héja, E. (2011) Multilingual term extraction from parallel corpora - A methodology for the automatic extraction of verbal structures and their translation equivalents. In: Fóris Ágota (Ed.) *Magyar Terminológia* 4 (2), p. 226-237. Akadémiai Kiadó

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

Wilcock, G. and K. Jokinen (2011). Adding Speech to a Robotics Simulator. *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop (IWSDS 2011)*, Granada, Spain, pp 371-376.