Finno-Ugric Digital Natives – Linguistic Support for Finno-Ugric Digital Communities

Kristiina Jokinen[1,2], Graham Wilcock[1,3], Tamás Varadi[4], Eszter Simon[4]
[1]University of Helsinki, [2]University of Tartu, [3]CDM Interact Oy
[4]Research Institute of Linguistics, Hungarian Academy of Sciences

Collaboration project between Academy of Finland and Hungarian Academy of Sciences

- Support digital visibility and viability of minor Finno-Ugric languages in the current globalising world
- Develop tools and resources for automatic language processing
- Generate online content on the target languages
- Experiment with interactive applications

Sami languages

- The nine Sami languages belong to the Finno-Ugric language family
- Spoken in the northern parts of Finland, Norway and Sweden and in the Kola Peninsula in Russia
- North Sami (*davvisámegiella*) is the biggest with ca 20 000 speakers
- Mainly morphophonetic variation, syntactic differences fairly small
- e.g. in Eastern dialects:
  - the vowel system contains three low vowels /ä/:/a/:/â/ Western dialects have only two /ä/:/a/
  - diphtongs become open:
    miessi /miessii/ ' fawn' : geassi /kiässii/: /kiessii/ 'summer'
    luohti /luohtii/ 'a yoik' : goahti /kuähtii/: kuohtii/ 'teepee, trad.dwelling'
  - /k/ in between vowels disappeares and /p/ changes into /v/
    johka /joohka/ : joga /joo.a/ 'river'
    lohpi /lohpii/ : lobi /loovii/ 'permission'

Due to active and strong revitalization efforts, through legislation and "language nests", the languages have gained an official status in the Nordic countries

Multimodal Interaction – SamiTalk

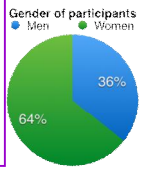
- Human-Robot Application for navigating information from Wikipedia
- Laughter and body posture recognition

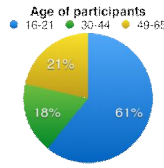
Robot: Man fáttá birra don háliidat gullat? Soames mu favorihta fáttát leat Sámegiella, Roavvenjárga, Mohtorgielká, Tiina Sanila-Aikio, Manga.
What topic do you want to hear about? Some of my favourites are: Sami language, Rovaniemi, Snowmobile, Tiina Sanila-Aikio, Manga.
User: Sámegiella (Sami language)
Robot: Sámegiela dilli Norgga, Ruota, Suoma ja Ruošša bealde lea earálágan.
The Sami language situation is different in Norway, Sweden, Finland and Russia
User: Dilli Suomas (The situation in Finland.)
Robot: Bealli Suoma 9350 sámis máhtet sámegiela. Suomas hállojuvvo davvi-sámegiella, anárašgiella ja nuortalašgiella, main anárašgiella dušše Suomas. …

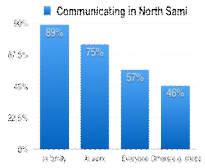
DigiSami Data Collection

- North Sami speech corpus: formal reading and free conversation
- Collected in 3 villages in Finland: Utsjoki, Inari and Ivalo, and in 2 villages in Norway: Kautokeino and Karasjoki
- Participants bilingual (Finnish or Norwegian), most had lived their life in the Sami area
- Data transcribed and used to study spoken interaction (speaker identification, multimodal conversation analysis, laughing)
- Available: https://goo.gl/SW9Goz

Dialects	Read speech		Conversational speech	
	#Spk	Duration	#Spk	Duration
Kautokeino	4	1.03	-	-
Karasjoki	6	0.72	6(1)	1.5
Ivalo	6	0.72	7(1)	0.72
Utsjoki	5	1.07	6(1)	1.03
Inari	4	0.73	-	-
Total	**25**	**3.26**	**19**	**4.28**







Dictionary Creation

Create bilingual dictionaries for six small Finno-Ugric languages (Udmurt, Komi-Permyak, Komi-Zyrian, Hill Mari, Meadow Mari and Northern Saami) paired with four major languages which are important for these communities (English, Finnish, Hungarian, Russian)

RESULTS FOR THE METHODS

method	all (#)	useful (%)
WikiTitle	2,989	99.33
W2D ext	921	98.91
W2D tri	11,714	93.51
KDE4	8,401	89.03

RESULTS FOR THE MERGED DICTIONARIES

lang pair	all (#)	useful (%)
sme–eng	6,042	92.29
sme–fin	7,100	91.44
sme–hun	4,969	90.72
sme–rus	4,373	95.95

- Jokinen, K., Hiovain, K., Laxström, N., Rauhala, I., Wilcock, G. DigiSami and Digital Natives: Interaction Technology for the North Sami language. *IWSDS*, 2016
- Jokinen, K. Open-domain Interaction and Online Content in the Sami Language. *LREC*, 2014
- Jokinen, K., Wilcock, G. Community-based Resource Building and Data Collection. *SLTU*, 2014
- Jokinen, K., NgoTrong, T., Hautamäki, V. Variation in Spoken North Sami Language, *Interspeech 2016*
- Benyeda, I., Koczka, P., Váradi, T. Creating seed lexicons for under-resourced languages. *GLOBALEX 2016*
- Simon, E., Benyeda, I., Koczka, P., Ludányi, Zs. Automatic creation of bilingual dictionaries for Finno-Ugric languages. *The 1st International Workshop on Computational Linguistics for Uralic Languages*, 2015
- Wilcock, G., Laxström, N., Leinonen, J., Smit, P., Kurimo, M., Jokinen, K. Towards SamiTalk: a Sami-speaking robot linked to Sami Wikipedia. *IWSDS*, 2016