



UNIVERSITY OF TARTU

Institute of Computer Science

TARTU^ULD

Large Models, Small Data

Mark Fišel

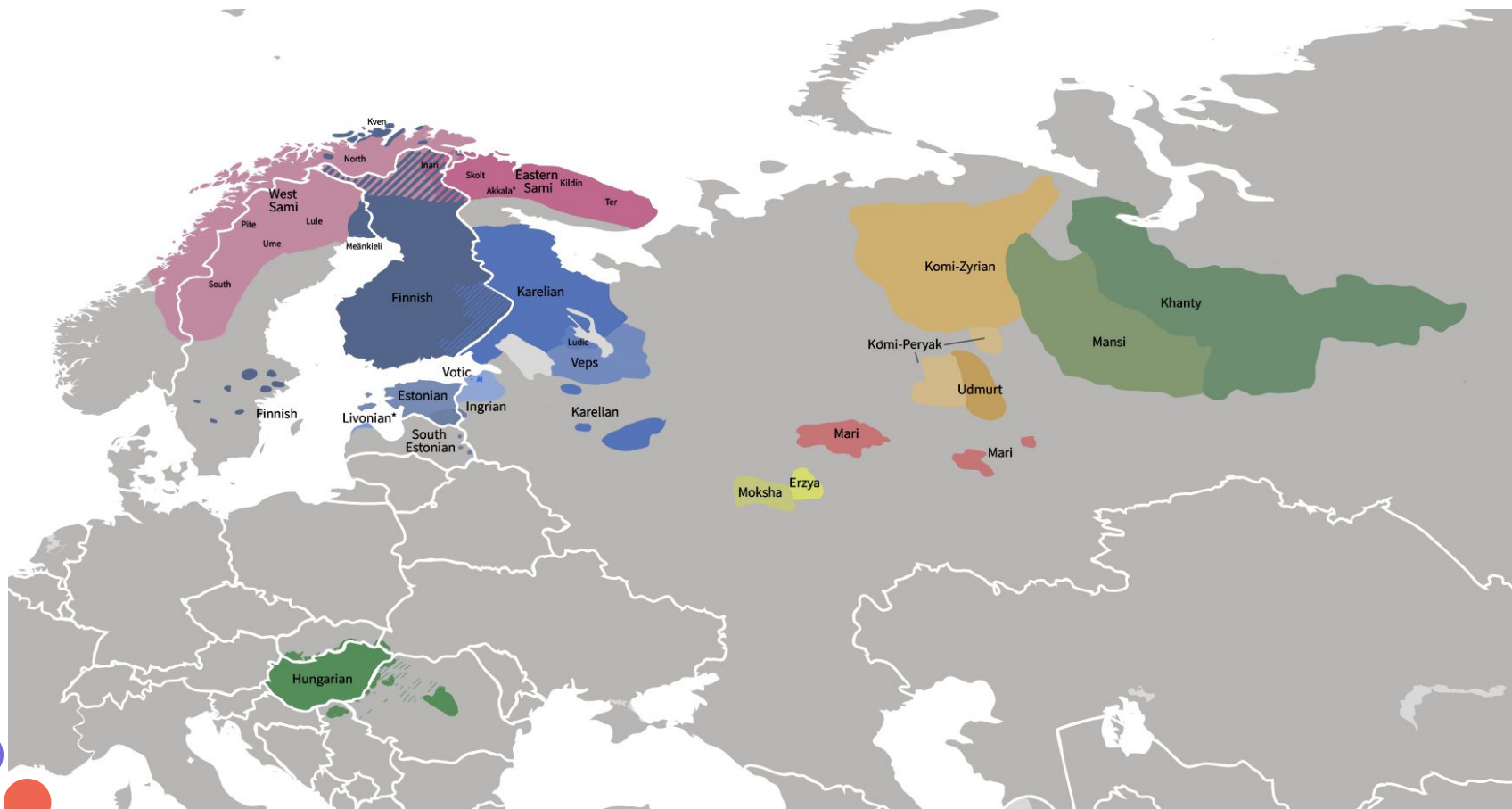
University of Tartu, Estonia

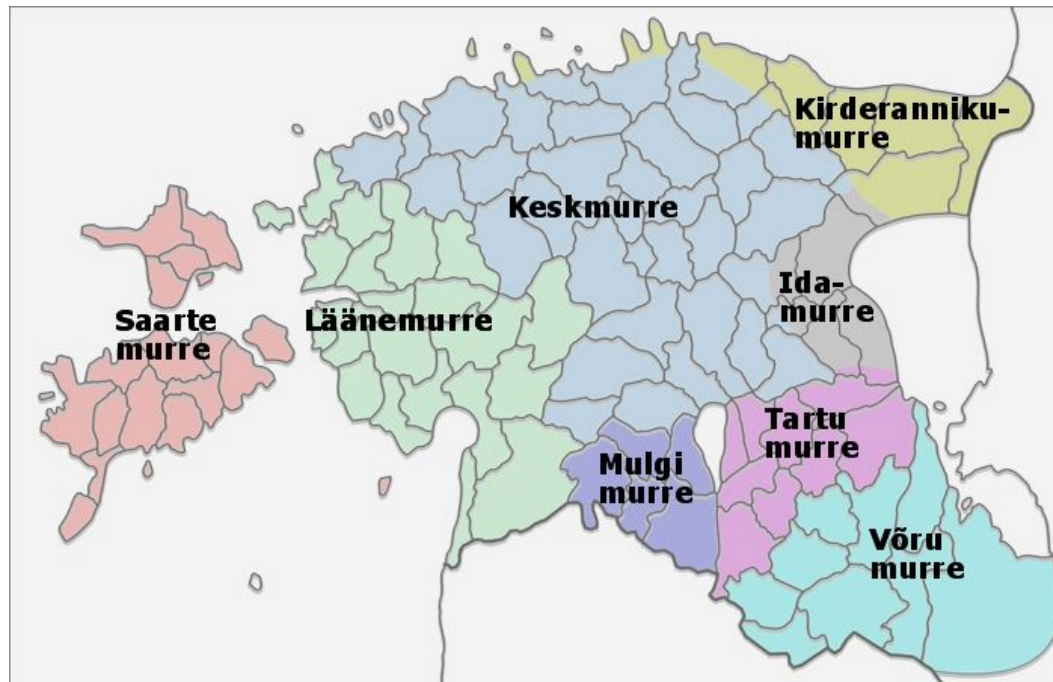
DigiTS seminar 2025, Tartu, Estonia

What we do and why (1)

- Everyone should be able to interface with technology **in their native language**
 - regardless of whether it is considered a language, (sub-)dialect, pidgin, etc.
- **AI should support languages**, not speed up their extinction









So far

2021: Võro - N. Sámi - S. Sámi - Estonian - Finnish MT



So far

2021: Võro - N. Sámi - S. Sámi - Estonian - Finnish MT

2022: Livonian - Estonian - Latvian - English MT
+ Inari/Skolt/Lule Sámi



So far

2021: Võro - N. Sámi - S. Sámi - Estonian - Finnish MT

2022: Livonian - Estonian - Latvian - English MT
+ Inari/Skolt/Lule Sámi

2023: + Veps, Karelian, Livvi, Ludic, Komi, Komi-Permyak, Udmurt,
Erzya, Moksha, Hill & Meadow Mari, Mansi, **Khanty** MT



So far

2021: Võro - N. Sámi - S. Sámi - Estonian - Finnish MT

2022: Livonian - Estonian - Latvian - English MT
+ Inari/Skolt/Lule Sámi

2023: + Veps, Karelian, Livvi, Ludic, Komi, Komi-Permyak, Udmurt,
Erzya, Moksha, Hill & Meadow Mari, Mansi, **Khanty** MT
+ Võro TTS
+ Estonian GEC



So far

2021: Võro - N. Sámi - S. Sámi - Estonian - Finnish MT

2022: Livonian - Estonian - Latvian - English MT
+ Inari/Skolt/Lule Sámi

2023: + Veps, Karelian, Livvi, Ludic, Komi, Komi-Permyak, Udmurt,
Erzya, Moksha, Hill & Meadow Mari, Mansi, Khanty MT
+ Võro TTS
+ Estonian GEC

2025 (TBD): + Ingrian, Votic, Ume/Kildin/Pite Sámi, Kven,... MT & LLM?



“Talk to me in Germanic”

Challenge #1: diversity in the data

- some dialects are more different than official languages
- we do not decide whether something is a language/dialect/..., but support it all anyway



“Talk to me in Germanic”

Challenge #1: diversity in the data

- some dialects are more different than official languages
- we do not decide whether something is a language/dialect/..., but support it all anyway
- lack of normalization in orthography + diverse sources: texts/translations, dictionaries, speech transcriptions, etc.
- ignoring it leads to a crazy mix when generating the output



Examples: orthography

Ingrian (Alamaluuga):

- ühekšä v~~o~~~~o~~ttä ei t~~e~~~~e~~ mi~~t~~ tää (Rozhanski & Markus)
- Meez jooksĩ moned kymmEnEd virssAd. (Muslimov)



Examples: orthography

Ingrian (Alamaluuga):

- ühekšä vootta ei tee miř tää (Rozhanski & Markus)
- Meez jookšĩ moned kymmEnEd virssAd. (Muslimov)

Votic (Std / Unk):

- Tšüzüb: "Ämme-üvä, missi sillõ on nii suurõd silmed?" (Books)
- isä ilmõz, poika purjõz. ahjo lämpiiB (EKI dictionary)



Examples: orthography

Ingrian (Alamaluuga):

- ühekšä vootta ei tee miř tää (Rozhanski & Markus)
- Meez jookšĩ moned kymmEnEd virssAd. (Muslimov)

Votic (Std / Unk):

- Tšüzüb: "Ämme-üvä, missi sillõ on nii suurõd silmed?" (Books)
- isä ilmõz, poika purjõz. ahjo lämpiiB (EKI dictionary)

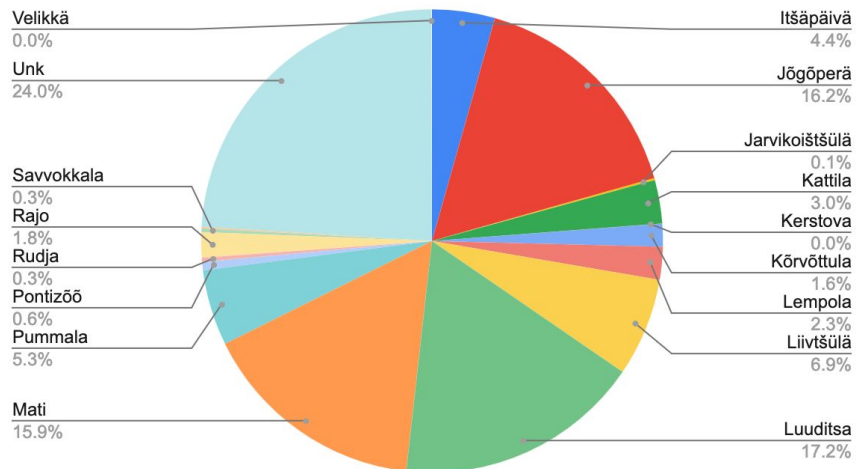
Võro:

- Panõq kirja niipall'o anõkduutõ (Sõnaq)
- ku saat hindäle fänni - tuu om su fänn kõgõs elos. (Umaleht)

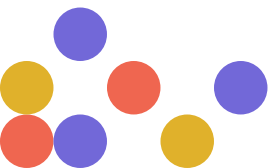
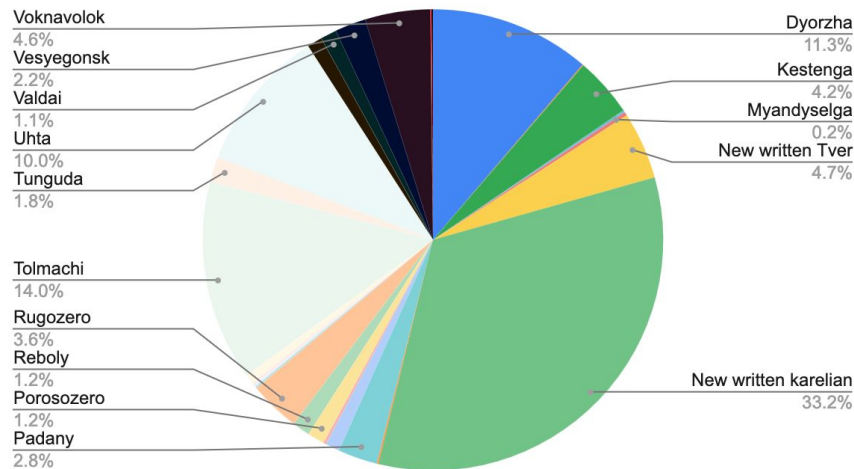


Dialects

Votic



Proper Karelian



“Google Translate said so”

Challenge #2: reliability

- publishing online tools needs to be done with care
 - AI (and human) output is never 100% correct
 - critical thinking needed
 - but not all users are knowledgeable in these languages!



“Google Translate said so”

Challenge #2: reliability

- publishing online tools needs to be done with care
 - AI (and human) output is never 100% correct
 - critical thinking needed
 - but not all users are knowledgeable in these languages!
- machine translation can be used to tackle the low resources
 - synthetic data
 - also heavily depends on the reliability of the output
 - need to predict the quality / confidence / uncertainty



Low-quality better than no data

Challenge #3: low-resource and diverse

- available data is small, diverse and includes different dialects, orthographies as well as originates from dictionaries, speech transcriptions, web, chats, etc.
 - but it has to be used, since that's all we have



Low-quality better than no data

Challenge #3: low-resource and diverse

- available data is small, diverse and includes different dialects, orthographies as well as originates from dictionaries, speech transcriptions, web, chats, etc.
 - but it has to be used, since that's all we have
- normalizing it orthographically and stylistically is challenging and can pollute the data further, if not done carefully



Examples: data sources

Dictionaries:

- kanal i kukol on tšivi, mikä java**В** →
“kanal ja kukel on kivi (= liivapugu), mis jahvatab (= peenendab toitu)”
- эрд; кушин → lage põld &v; väli
- кече уже пеш күшнө → päike on [juba] suures lõunas &v; suures kõrges



Examples: data sources

Dictionaries:

- kanal i kukol on tšivi, mikä java**В** →
“kanal ja kukel on kivi (= liivapugu), mis jahvatab (= peenendab toitu)”
- эрд; кушин → lage põld &v; väli
- кече уже пеш күшнө → päike on [juba] suures lõunas &v; suures kõrges

Speech transcriptions:

- <ka> kold poiķ kaa i meil oli tüttöi
- ??? en puuttuund
- ei jädedä še kaig varaššet taa män nöö kuhu lee toiš šee pöölee ja →
[They] leave nothing, [they] steal everything, [everything] goes somewhere to another side [= abroad].



Our approach

- **Multilinguality**
 - cross-lingual transfer-learning: languages “support” each other



Our approach

- **Multilinguality**
 - cross-lingual transfer-learning: languages “support” each other
- **Data sources**
 - **no normalization**: train the models with raw data
 - but **inform it** about the dialect, orthography, data source, etc.

“translate this from Ingrian, Soikkola, speech transcription to Estonian, speech transcription: ...”



Our approach

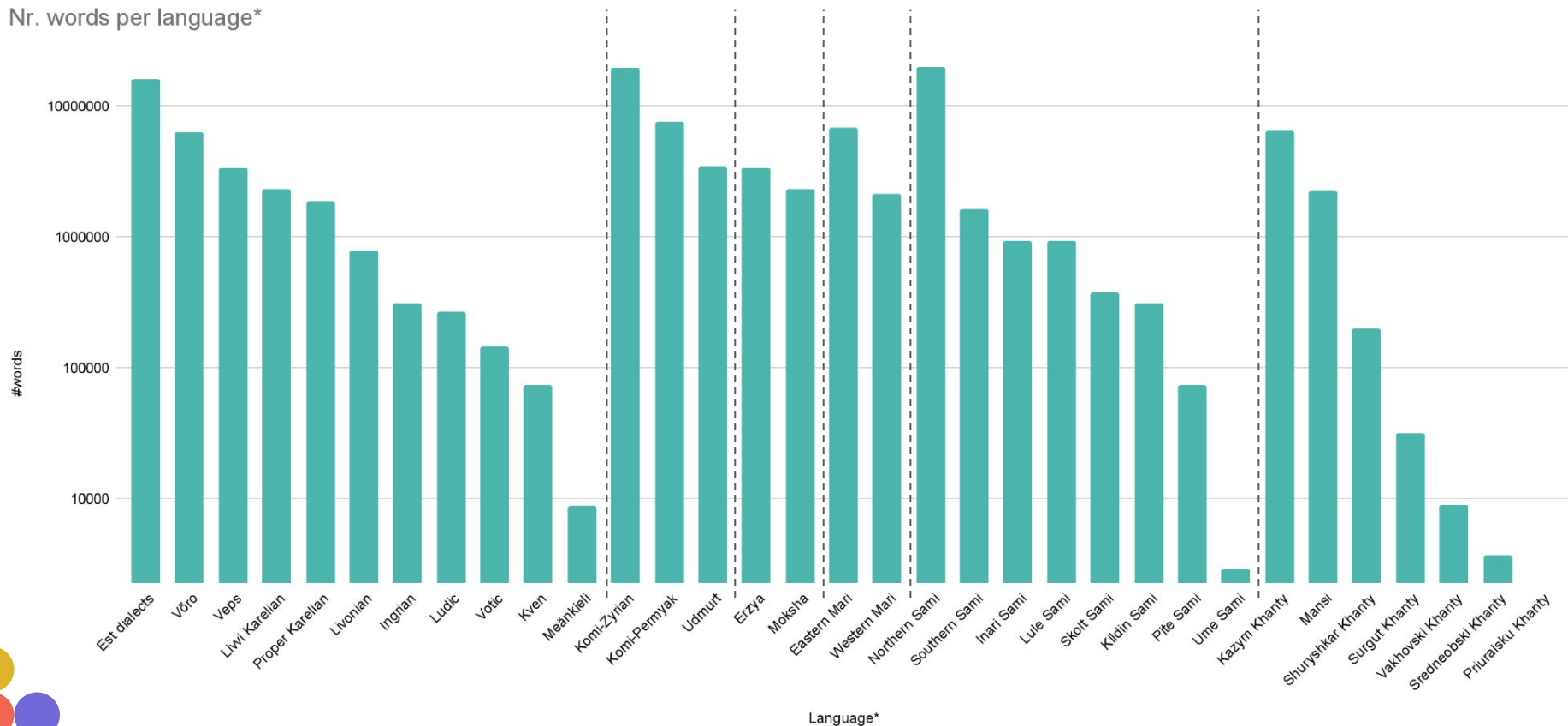
- **Multilinguality**
 - cross-lingual transfer-learning: languages “support” each other
- **Data sources**
 - no normalization: train the models with raw data
 - but **inform it** about the dialect, orthography, data source, etc.

“translate this from Ingrian, Soikkola, speech transcription to Estonian, speech transcription: ...”
- **Current focus: machine translation**
 - useful for generating data for LLMs and elsewhere



Amount of text data

Nr. words per language*



Translation data

	Võro	Seto	Kihnu	Mulgi	Rannakeel	Hiiu	Votic	Ingrian	Livonian	Veps	Karelian Proper	Livvic	Ludian	Meänkieli	Kven	Komi-Zyrian	Komi-Permyak	Udmurt	Erzya	Moksha	Meadow Mari	Hill Mari	Kildin Sami	Pite Sami	Ume Sami	S. Sami	N. Sami	Lule Sami	Inari Sami	Skolt Sami	Mansi	W.Khanty	E.Khanty	
Estonian	347,493.0	283,546.0	65,982.0	26,896.0	9,888.0	9,096.0	97,299.0	1,785.0	134,831.0	62,176.0	56,851.0	60,057.0				154,903.0	58,770.0	434,256.0	372,251.0	72,288.0	392,721.0	62,109.0									3,546.0	16,544.0		
English	102.0						2,510.0	41,536.0	62,368.0	116,055.0	97,489.0	99,483.0			2,952.0	103,172.0	86,212.0	343,760.0	123,114.0	137,261.0	113,041.0	88,463.0	1,933.0	23,480.0			10,804.0	9.0	2,368.0	1,540.0	8,447.0	26,190.0		
Russian								46,079.0		600,188.0	610,678.0	584,576.0	213,586.0		535.0	16,992,104	250,047.0	1,160,882	385,521.0	395,784.0	3,027,703	247,037.0	173,866.0				461.0				24,876.0	99,336.0	2,908.0	
Finnish										115,873.0	99,680.0	100,270.0	804.0		36.0	102,989.0	86,095.0	338,472.0	122,954.0	137,107.0	108,341.0	88,307.0	597.0	118.0		29,079.0	419,386.0		137,156.0	95,573.0	7,550.0	23,666.0		
Hungarian										116,010.0	97,452.0	99,453.0				103,162.0	86,182.0	338,894.0	122,986.0	137,268.0	108,468.0	88,433.0		74.0							7,579.0	23,726.0		
Latvian									108,844.0	96,668.0	96,822.0	92,600.0				78,799.0	86,084.0	93,980.0	92,080.0	97,101.0	89,384.0	88,324.0									6,993.0	7,557.0		
Norwegian										116,004.0	97,462.0	99,447.0			55,718.0	103,023.0	86,147.0	338,561.0	123,042.0	137,229.0	108,402.0	88,404.0			3,957.0		247,688.0	2,974,036.0	13,392.0		2,356.0	7,579.0	23,704.0	
Swedish											1675	925	804	321	2361									18833	2260	2215	4607	2256						
German																							597	2961										



Results: scores (chrf++)

input ↓ \ output →	English	Estonian	Finnish	Hungarian	Latvian	Norwegian	Russian	Võro	Proper Karelian	Livvi	Ludian	Veps	Livonian	Erzya	Moksha	Hill Mari	Meadow Mari	Komi Zyrian	Udmurt	Mansi	Kazym Khanty
English		57.2	56.8	56.1	58.1	59.2	53.9	35.4	44.8	37.6	31.5	37.2	30.5	32.7	37	25.8	41.9	41.7	32.8	15.6	10.8
Estonian	61.9		52.5	51	52.4	47.1	48.3	43.7	43.9	36.9	31.2	34.3	33.2	31.8	40.5	25.7	41.4	41.6	35.9	15.3	10.3
Finnish	59	52.4		49.5	51	48.3	46.6	34.6	52.6	39.3	34.2	36.1	28.2	31.5	35.2	25.9	39.8	38.8	33.2	14.8	11
Hungarian	59.8	52.4	50.2		51.3	47.2	47.3	34.6	42.3	36.1	30	33.3	29.2	27	35.8	25	40.2	39.6	30.4	13.6	10.4
Latvian	61	53.7	52.2	50.7		49.6	52.6	34.6	43	37.4	31.6	35.9	29.6	31.1	36.1	27.4	43	41.8	36.4	13.5	10.6
Norwegian	66.9	51.7	50.6	50.3	50.7		48.9	34.4	42.7	36.8	29.8	34.6	29.7	31.6	36	25.4	40	38.9	33.9	13.8	10.3
Russian	59.6	50.7	45	47.9	53	47.5		33.8	44.2	41.9	33.4	42.8	28.6	43	39.1	37.6	51.2	49.8	47.3	17.8	11.4
Võro	53.2	68.9	48.7	46.3	46.4	43.5	43.2		40.1	35	30.1	33.1	32.7	32.6	31.5	26.6	37.1	38.4	32.8	13	9.2
Proper Karelian	52.8	46.8	59.8	46.6	46.8	43.7	45.9	31.7		40.6	31.9	37.1	27.6	31.9	33.2	28.8	36.7	38.3	33.7	13.7	11.1
Livvi	46.3	42.7	48.8	41.2	43.5	39.5	48	29.5	38.5		34.5	38.8	26	35.6	32.4	29.6	38.6	39.6	35.2	15.9	10.7
Ludian	45.3	44.3	49.8	40	41.6	39.2	42.6	30.8	39.8	40.1		36.6	26.6	34.8	33.4	29.4	37.1	38.9	33.8	15.1	11.1
Veps	44	40.9	43.5	39.7	42.1	37.3	45.9	29	35.1	38.4	33.3		26.5	34	32.2	30.2	39.1	40.1	34.2	15.4	11.5
Livonian	49.5	52.3	43.5	39.9	43.5	41.6	39.5	35.3	37.9	33.4	29.4	32.7		27.5	33.6	25.7	33.2	35.7	30	13.5	10.9
Erzya	46.1	40.5	42.1	39.8	42.5	39.3	52.3	27.2	36.9	35.2	27.6	35.8	25.2		36.1	31	43.1	42	39.3	16.8	11.4
Moksha	43.4	45.9	40.8	39	41.8	38.1	38.5	31.8	35.1	31.6	27.3	32.3	25.2	34.6		28	36	37.6	33.5	16.1	12.5
Hill Mari	46.5	40.1	43.1	41.1	44.3	40.4	50.2	28.2	35.5	34.6	25.7	35	24.9	34.5	32.1		45	41	39.3	19.1	11.5
Meadow Mari	48.4	41.6	43.8	42.1	45.4	41.5	53.1	29.5	37.3	34.9	26	35.4	25.9	36.8	35.3	44.4		42.4	42.4	18.5	12
Komi-Zyrian	49.5	45.1	44	42.6	45.8	41.3	53.3	30.2	37.3	37.1	29.2	36.8	26.9	38.6	35.8	32.3	43.1		40	16.8	11.4
Udmurt	47.8	43.7	43.1	41.9	44.2	40.2	54.9	29.9	37.4	37	28.7	36.5	25.3	38.9	35.3	37.3	45.5	44.8		19.1	12.3
Mansi	36.7	31.6	32.6	33	33.4	32	35.7	23.2	28.5	27	22.6	27.3	20.6	27	25.4	26.4	33.3	31.2	29.8		11.8
Kazym Khanty	33.2	29.8	31.3	29.9	30.2	29.2	31.5	20.8	25	26.3	18	24.3	19.1	21.2	24.4	21	29.6	30.2	26.4	15.7	
Avg gen score	50.5	46.6	46.1	43.4	45.4	42.3	46.6	31.4	38.9	35.9	29.3	34.8	27.1	32.8	34.0	29.2	39.7	39.6	35.0	15.7	11.1

Results: qualitative

To	Meaning						Language			chrF	data
	Prf	Und	Part	Poor	Unusbl	CntSay	Cor	WrDi	WrLng		
from English											
Kildin Sámi	22.2	16.7	20	27.8	13.3	0	96.7	1.1	2.2	–	
Pite Sámi (60)	1.7	23.3	16.7	16.7	40	1.7	76.7	20	3.3	–	
Skolt Sámi (30)	6.7	13.3	23.3	26.7	30	0	90	0	10	–	
Northern Sámi (30)	40	46.7	10	3.3	0	0	100	0	0	–	
Livonian	20	25.6	18.9	14.4	13.3	7.8	65.6	23.3	11.1	30.5	
Livonian	11.1	24.4	32.2	16.7	15.6	0	85.6	4.4	10	30.5	
Livonian (30)	6.7	10	50	30	3.3	0	66.7	33.3	0	30.5	
from Estonian											
Võro	26.7	37.8	20	10	2.2	3.3	92.2	5.6	2.2	43.7	
Livonian	11.1	23.3	41.1	24.4	0	0	100	0	0	33.2	
Livonian	16.7	36.7	24.4	13.3	7.8	1.1	61.1	33.3	5.6	33.2	
Livonian	11.1	36.7	35.6	15.6	1.1	0	76.7	23.3	0	33.2	
Livonian (30)	53.3	20	3.3	23.3	0	0	90	10	0	33.2	
from Russian											
Komi Zyrian	34.4	54.4	7.8	1.1	2.2	0	92.2	0	7.8	49.8	
Livvi Karelian	23.3	76.7	0	0	0	0	100	0	0	41.9	
Izhorian	2.2	11.1	25.6	28.9	32.2	0	65.6	20	14.4	–	
Votic	1.1	17.8	43.3	25.6	12.2	0	16.7	81.1	2.2	–	
Kazym Khanty (30)	0	26.7	6.7	36.7	30	0	83.3	16.7	0	11.4	

People:

- Data:
 - Heiki-Jaan Kaalep
 - Tarmo Vaino
 - Aleksei Ivanov
 - Britt-Kathleen Mere
 - Kaire Koljal
 - Annely Liivas
- Machine translation:
 - Lisa Yankovskaya
 - Dmytro Pashchenko
- LLMs
 - Hele-Andra Kuulmets
 - Taido Purason
- Speech processing
 - Liisa Rätsep
 - Rasmus Lellep
- Grammatical error correction
 - Agnes Luhtaru
 - Martin Vainikko



Partners & data providers

- Eesti keele instituut
- Võro instituut
- LU Līvõd institūt
- Itä-suomen yliopisto
- Norges arktiske universitet
- New Amigos
- Miikul Pahomov
- Michael Rießler
- Joshua Wilbur
- Elena Markus
- Fedor Rozhanskiy
- Mehmet Muslimov
- Irina Moldanova
- Tuuli Tuisk
- Marili Tomingas
- Miina Norvik
- Valts Ernštreits
- ...



Supporters

- Estonian Research Council
- National Programme “Estonian Language Technology”
- Institute of the Estonian Language
- UniTartu HPC, LUMI supercomputer



Tienū!