SMUGRI-MT:

Machine Translation for Low-Resource Finno-Ugric Languages

Lisa Yankovskaya, PhD
Tartu NLP, University of Tartu
September 30, 2024

TARTUNLP

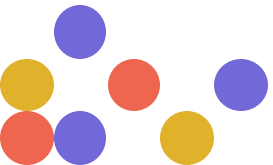
Soome-ugri keeled

Suopmelaš-ugralaš gielat

SMUGRI: Finno-Ugric NLP

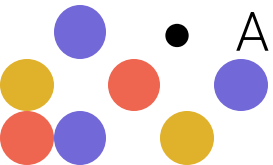
Суоми-Угрань кяльхне

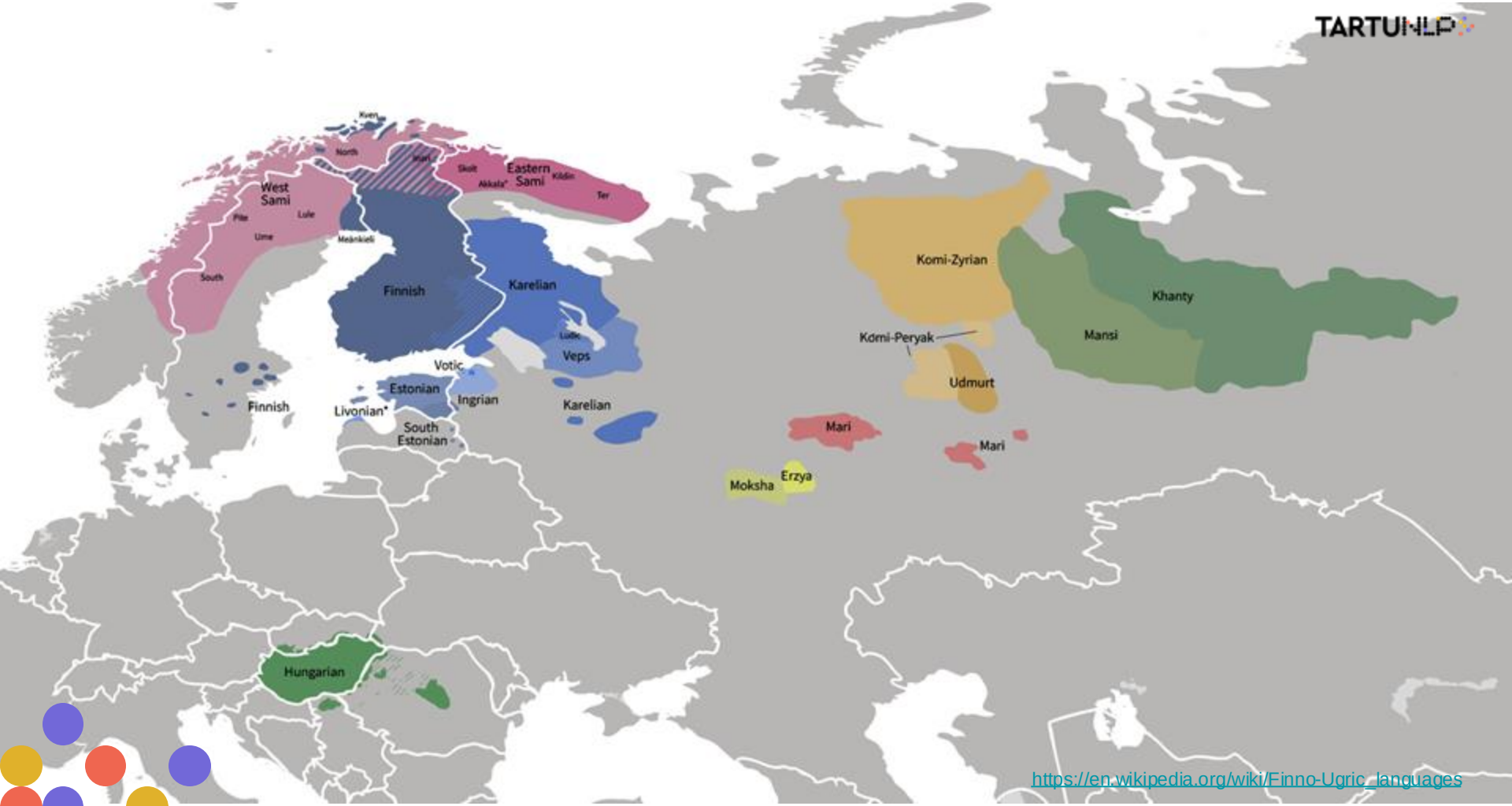
Šuomelais-ugrilaiset kielet



Outline

- Finno-Ugric Languages: data
- Results: what we have done so far
- Challenges
- Next Steps
- Are the members of community users of MT systems?




TARTUNLP
West Sami
North
Pite
Lule
Ume
South
Meänkieli
Finnish
Skolt
Akkala*
Eastern Sami
Kildin
Ter
Karelian
Ludic
Veps
Votic
Estonian
Ingrian
Livonian*
South Estonian
Finnish
Karelian
Komi-Zyrian
Khanty
Komi-Peryak
Mansi
Udmurt
Mari
Mari
Moksha
Erzya
Hungarian
https://en.wikipedia.org/wiki/Finno-Ugric_languages

TARTUNLP

#speakers / #resources

MT progress:

2021: Võro - Estonian + Northern&Southern Sámi - Finnish MT

MT progress:

2021: Võro - Estonian + Northern&Southern Sámi - Finnish MT

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

MT progress:

2021: Võro - Estonian + Northern&Southern Sámi - Finnish MT

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

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

MT progress:

2021: Võro - Estonian + Northern&Southern Sámi - Finnish MT

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

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

2024: **Paragraph-level MT** (Veps, Proper Karelian, Ludic, Livvi)

TARTUNLP

Coverage:

	Google		m2m100	NLLB	MADLAD-400	MMS (ASR)	MMS (TTS)
Hungarian	✓	✓	✓	✓	✓	✓	✓
Finnish	✓	✓	✓	✓	✓	✓	✓
Estonian	✓	✓	✓	✓	✓	✓	✓

	Google		m2m100	NLLB	MADLAD-400	MMS (ASR)	MMS (TTS)
Hungarian	✓	✓	✓	✓	✓	✓	✓
Finnish	✓	✓	✓	✓	✓	✓	✓
Estonian	✓	✓	✓	✓	✓	✓	✓
Komi	✓	✗	✗	✗	✓	✓	✓
Udmurt	✓	✓	✗	✗	✓	✓	✓
Erzya	✗	✗	✗	✗	✓	✓	✓
Meadow Mari	✓	✓	✗	✗	✓	✓	✓
Hill Mari	✗	✓	✗	✗	✓	✗	✗
Komi-Permyak	✗	✗	✗	✗	✓	✗	✗
Karelian	✗	✗	✗	✗	✗	✓	✓
Khanty	✗	✗	✗	✗	✗	✓	✓
Moksha	✗	✗	✗	✗	✓	✓	✗
Northern Sami	✓	✗	✗	✗	✓	✗	✗
Inari Sami	✗	✗	✗	✗	✗	✗	✗
Livvi Karelian	✗	✗	✗	✗	✗	✗	✗
Votic	✗	✗	✗	✗	✗	✓	✗
Võro	✗	✗	✗	✗	✗	✗	✗
Mansi	✗	✗	✗	✗	✗	✗	✗
Veps	✗	✗	✗	✗	✗	✗	✗

Coverage:

← Some languages covered, however:

- Many more missing: Livonian, Ludian, Lule/Southern/Pite/Skolt/Kildin/… Sami (between 10 and ~75k speakers), etc.

Though some previous work exists:

- e.g. FST morphology, translation, etc for Sami and other smugri languages

Results: benchmark

- FLORES-200: 1k sentences, translated into 200 languages by Meta
  - incl. Estonian, Finnish & Hungarian

We hired human translators to translate 250 sentences into

- Komi
- Udmurt
- Hill Mari
- Meadow Mari
- Proper Karelian
- Ludian
- Livvi Karelian
- Veps
- Võro
- Erzya
- Moksha
- Livonian
- Mansi

Results

Output lang	Average chrF++
Russian	48.7
English	50.2
Estonian	46.3
Finnish	43.7
Hungarian	41.5
Latvian	43.1
Norwegian	40.4

Output lang	Average chrF++
Komi	39.6
Livonian	29.4
Moksha	34.2
M.Mari	39.9
Mansi	23.4
H.Mari	37.4
Erzya	33.7

Output lang	Average chrF++
L.Karelian	31.0
Udmurt	36.6
P.Karelian	41.6
Võro	39.9
Veps	34.6
Ludian	28.3

Results

Output lang	Average chrF++
Russian	48.7
English	50.2
Estonian	46.3
Finnish	43.7
Hungarian	41.5
Latvian	43.1
Norwegian	40.4

Output lang	Average chrF++
Komi	39.6
Livonian	29.4
Moksha	34.2
M.Mari	39.9
Mansi	23.4
H.Mari	37.4
Erzya	33.7

Output lang	Average chrF++
L.Karelian	31.0
Udmurt	36.6
P.Karelian	41.6
Võro	39.9
Veps	34.6
Ludian	28.3

Results

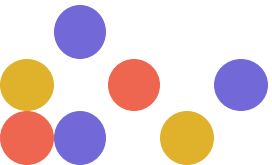
Output lang	Average chrF++
Russian	48.7
English	50.2
Estonian	46.3
Finnish	43.7
Hungarian	41.5
Latvian	43.1
Norwegian	40.4

Output lang	Average chrF++
Komi	39.6
Livonian	29.4
Moksha	34.2
M.Mari	39.9
Mansi	23.4
H.Mari	37.4
Erzya	33.7

Output lang	Average chrF++
L.Karelian	31.0
Udmurt	36.6
P.Karelian	41.6
Võro	39.9
Veps	34.6
Ludian	28.3

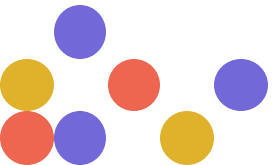
Challenges

- scarce resources



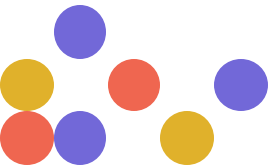
Challenges

- scarce resources → collect what we can, exploit multilinguality



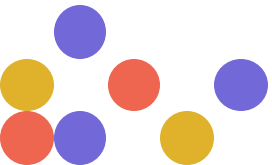
Challenges

- scarce resources → collect what we can, exploit multilinguality
- dialect / supradialect / language / ?



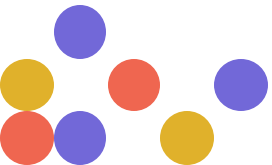
Challenges

- scarce resources → collect what we can, exploit multilinguality
- dialect / supradialect / language / ?
  - Collecting data



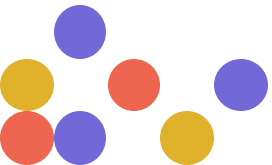
Challenges

- scarce resources → collect what we can, exploit multilinguality
- dialect / supradialect / language / ?
  - Collecting data → we try to differentiate



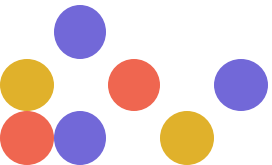
Challenges

- scarce resources → collect what we can, exploit multilinguality
- dialect / supradialect / language / ?
  - Collecting data → we try to differentiate
  - Separate translation direction?



Challenges

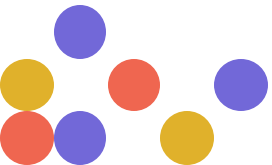
- scarce resources → collect what we can, exploit multilinguality
- dialect / supradialect / language / ?
  - Collecting data → we try to differentiate
  - Separate translation direction? →Yes, if we have enough resources and the (sub)dialects are distinct



Challenges

- scarce resources → collect what we can, exploit multilinguality

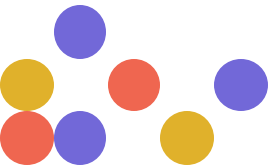
- dialect / supradialect / language / ?
  - Collecting data → we try to differentiate
  - Separate translation direction? →Yes, if we have enough resources and the (sub)dialects are distinct
  - During training?



Challenges

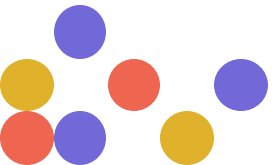
- scarce resources → collect what we can, exploit multilinguality

- dialect / supradialect / language / ?
  - Collecting data → we try to differentiate
  - Separate translation direction? →Yes, if we have enough resources and the (sub)dialects are distinct
  - During training? → we do not differentiate



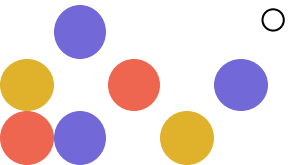
Challenges

- scarce resources → collect what we can, exploit multilinguality
- dialect / supradialect / language / ?
  - Collecting data → we try to differentiate
  - Separate translation direction? →Yes, if we have enough resources and the (sub)dialects are distinct
  - During training? → we do not differentiate
- not standardized writing / different alphabets

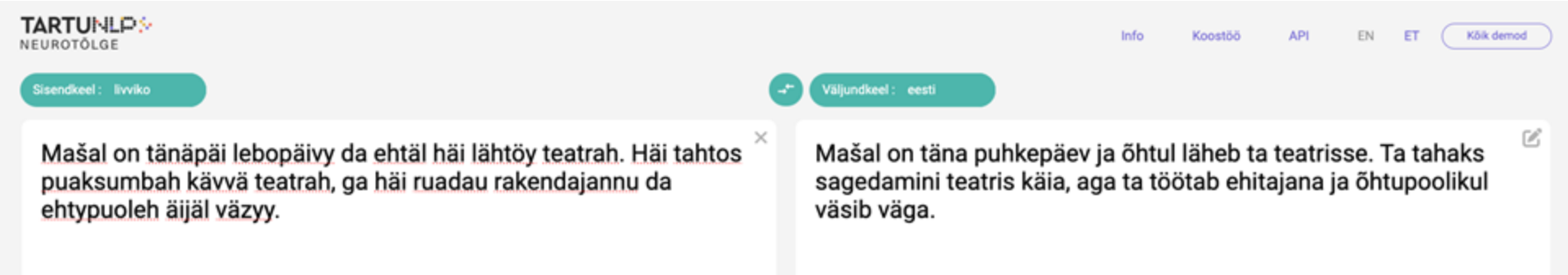


Challenges

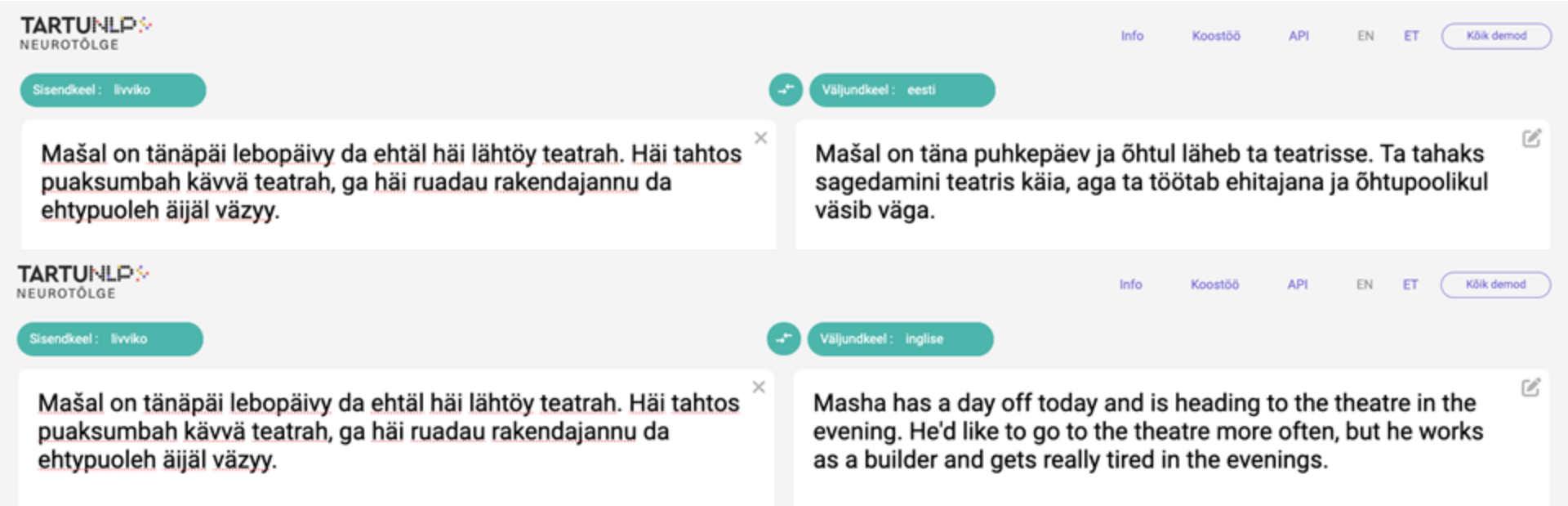
- scarce resources → collect what we can, exploit multilinguality
- dialect / supradialect / language / ?
  - Collecting data → we try to differentiate
  - Separate translation direction? →Yes, if we have enough resources and the (sub)dialects are distinct
  - During training? → we do not differentiate
- not standardized writing / different alphabets → filter, clean up / ignore
  - affects overall quality



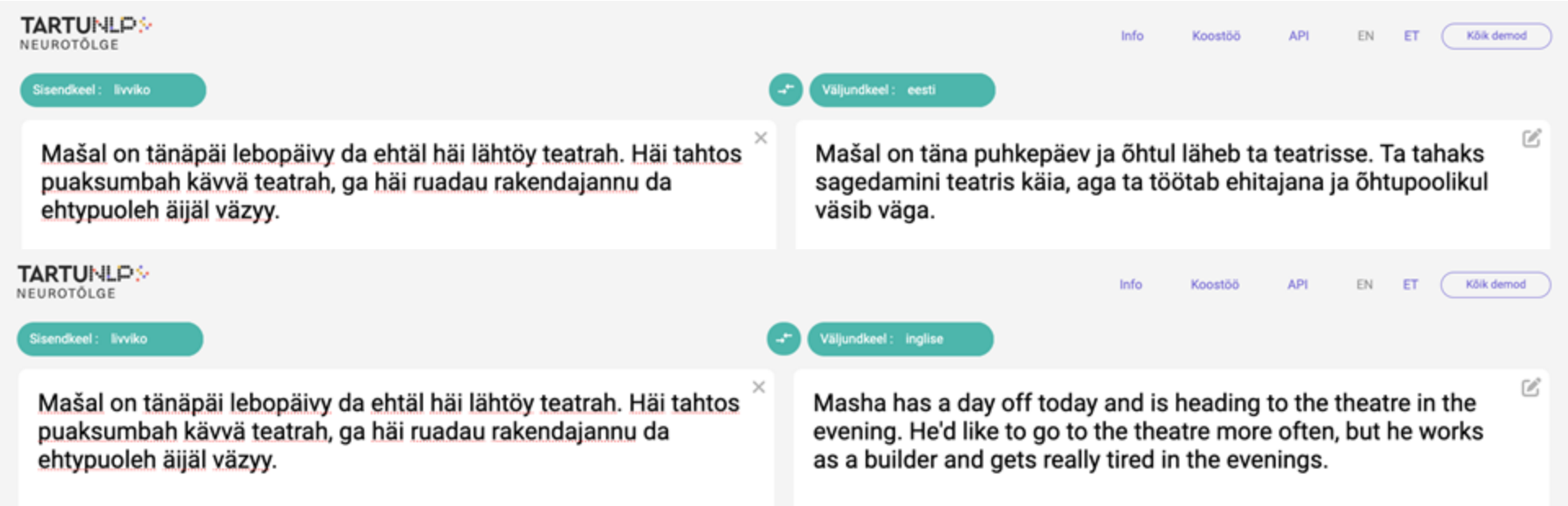
translate.ut.ee



translate.ut.ee



translate.ut.ee



Context-aware system →

Masha has a rest day today, and in the evening she goes to the theater. She would like to go to the theater more often, but she works as a builder and gets very tired in the evening.

Next steps

- paragraph-level MT or LLMs?
- add Pite Sami, Kildin Sami, Ingrian, Votic, more Estonian dialects and every other smugri language we can lay our hands on
- separate translate direction for each dialect/subdialect
- add translation quality estimation?

Is MT what is needed?

Is MT what is needed?

Wiechetek et al, LREC'2022: Unmasking the Myth of Effortless Big Data — Making an Open Source Multilingual Infrastructure and Building Language Resources from Scratch:

"Machine translation **into a minority language** is of no use when the output is unreliable and the language community bilingual, thus in a position to choose the majority language original instead. Translation **from the minority language** is of no use if there is no monolingual text to translate."

Is MT what is needed?

- Speakers?
- Learners?
- People who wants to know about their ancestors?
- Researchers?

Supporters

- Estonian Research Council
- National Programme “Estonian Language Technology”
- Institute of the Estonian Language
- Võru institute
- UL Livonian Institute
- University of Eastern Finland

TARTUNLP

Thank you!

huggingface.co/tartuNLP
github.com/tartunlp