



UNIVERSITY OF TARTU

Institute of Computer Science

TARTUMLP

Mašintõlge: kā töötab, plussid/mīnusi, mis kīeldkõks se strōdõb?

Mark Fišel, TÜ keeletehnoloogia professor

9.11.2024

Tutvustus₁

- 2006 -- 2011: PhD, masintõlge
- 2011 -- 2015: post-doc, masintõlge
- 2015 -- praegu: keeletehnoloogia õppetooli juhataja (sh masintõlge)



Tutvustus₂

- kõnesüntees (neurokone.ee)
- suured keelemudelid
- automaatne keeleline korrektuur
- ja masintõlge (neurotolge.ee)



Plaan

- kuidas töötab ja õpib tänapäeva masintõlge
- mis mõju on suurtel keelemudelitel masintõlkele
- eesti-läti-liivi masintõlke arendamine



Masintõlge: milleks?

- ~~selleks, et asendada inimtõlkijaid masinatega, rikkuda kõiki keeli ära ning vallutada maailma~~
- selleks, et inimtõlkijail oleks kaasaegne tööriist, mis võimaldab
 - kiirendada nende tööd
 - saada rutiinsemate tekstide tõlkimisest lahti
 - keskenduda keerulisematele juhtumitele, millega masin kunagi hakkama ei saa



Masintõlge

- kas ideaalseks tõlkimiseks peab tekstist **aru saama**? kas peab olema suuteline **mõelda, tunda**?



Masintõlge

- kas ideaalseks tõlkimiseks peab tekstist **aru saama**? kas peab olema suuteline **mõelda, tunda**? (Mark: **jah**)



Masintõlge

- kas ideaalseks tõlkimiseks peab tekstist **aru saama**? kas peab olema suuteline **mõelda, tunda**? (Mark: **jah**)
- kas arvutiprogrammid seda suudavad? **EI**
 - seega tegu on hoopis tõlkimise **imiteerimisega**



Masintõlge: enne

Kuigi masintõlge teeb palju naljakaid vigu,
võib see siiski praktikas kasulik olla



Masintõlge: enne

Kuigi masintõlge teeb palju naljakaid vigu,
võib see siiski praktikas kasulik olla

I know you are the dream girl of
millions of your fans



Ich weiß, dass Sie ein Traumädchen sind
von den Millionen Ihrer Ventilatoren.

Masintõlge: enne

Kuigi masintõlge teeb palju naljakaid vigu,
võib see siiski praktikas kasulik olla

Inglise ▼

Eesti ▼

I am taking a German course.

0/500

Tõlgi

Mängi

 Ma võtan Saksa hooaega.

 Ma sõidan Saksa kursusega.

 Ma võtan Saksa muidugi.

I know you are the dream girl of
millions of your fans



Masintõlge: nüüd

Kuigi võib tunduda, et masintõlge on lahendatud,
siiski vajab see veel arendamist



 Text Images Documents WebsitesDetect language Estonian Hebrew ▼↔ Latvian Latgalian Sami (North) ▼

Naine jõudis koju. Ta jättis auto
garaaži ning läks magama. ✕



59 / 5,000



Sieviete ieradās mājās. Viņš atstāja ☆
automašīnu garāžā un devās gulēt.



ChatGPT ▾

Tõlgi tekst "Naine jõudis koju. Ta jättis oma auto garaaži ning läks tuppa." eesti keelest inglise keelde.



The woman arrived home. She left her car in the garage and went inside.



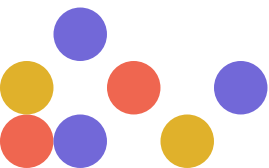
Message ChatGPT



ChatGPT can make mistakes. Check important info.



chat.openai.com





Tõlgi tekst "I need an extension for my visa. Tell me how can I apply for it" eesti keelde -- seejuures aga kohenda stiili nii, et tõlge kõlaks ilusasti, väga viisakalt ja ametlikult.



"Eelistan oma viisa pikendamist. Palun andke teada, kuidas saan selleks taotluse esitada."



Message ChatGPT



Suured keelemudelid

- + suudavad vastata küsimustele ja täita ülesandeid
- + “oskavad” mitut keelt
- + saavad hakkama ka ülesannetega, mida nad õppimise käigus ei ole näinud
 - o väga hea üldistamisvõime



Suured keelemudelid

- on aeglased ja nende loomine on kallis
- ei ole usaldusväärsed -- vahest on nende vastused valed
 - o kuid võivad kõlada kindlalt
- ei oska kõiki keeli võrdselt hästi



LLM töö:

“Anna palun vett, ma” → “tahan”

“Anna palun vett, ma tahan” → “juua”

“Anna palun vett, ma tahan juua” → “.”

“Anna palun vett, ma tahan juua.” → “⟨lõpp⟩”



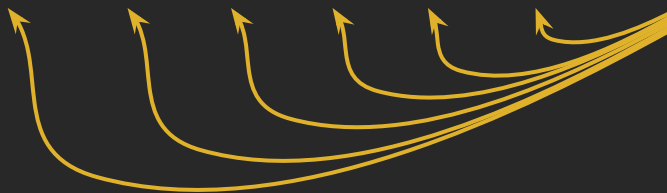
“Tähelepanu”:

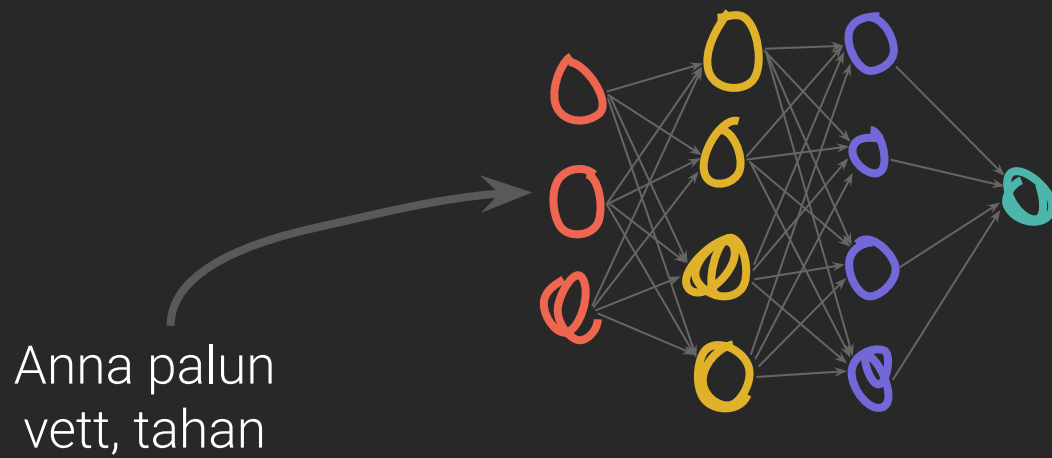
“Anna palun vett, ma” → “tahan”

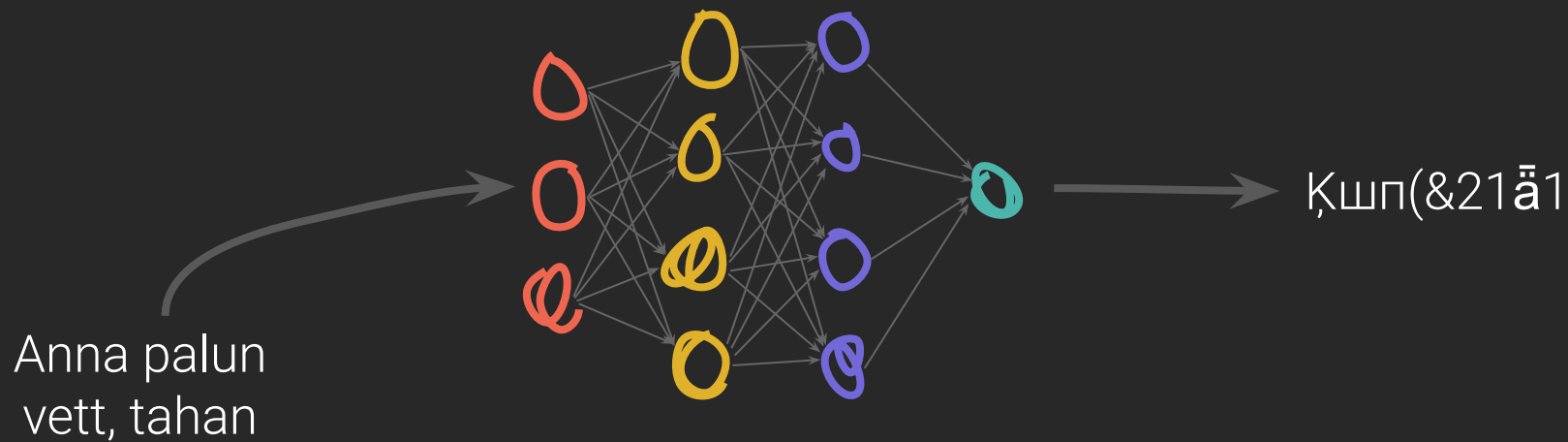
“Anna palun vett, ma tahan” → “juua”

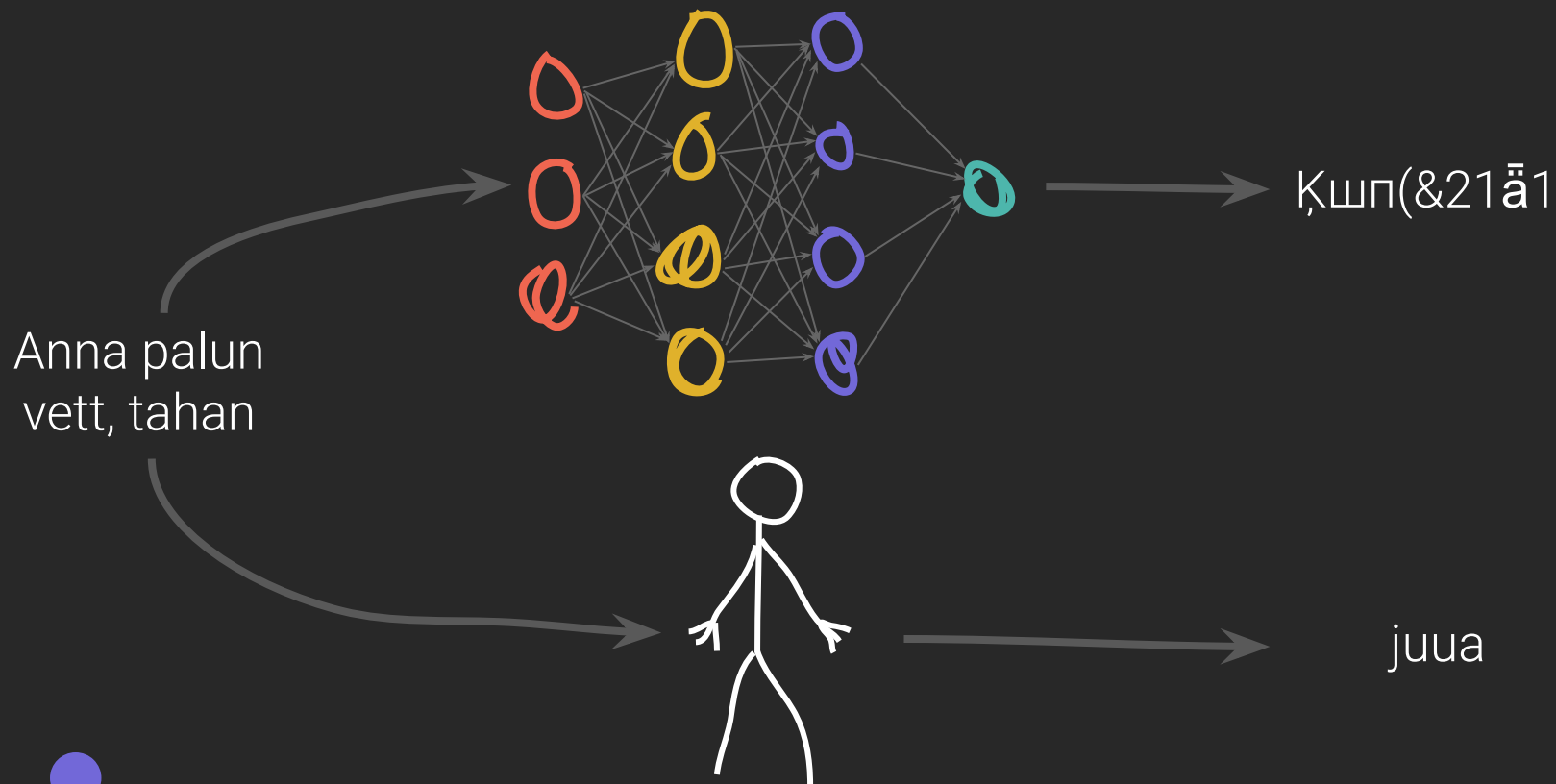
“Anna palun vett, ma tahan juua” → “.”

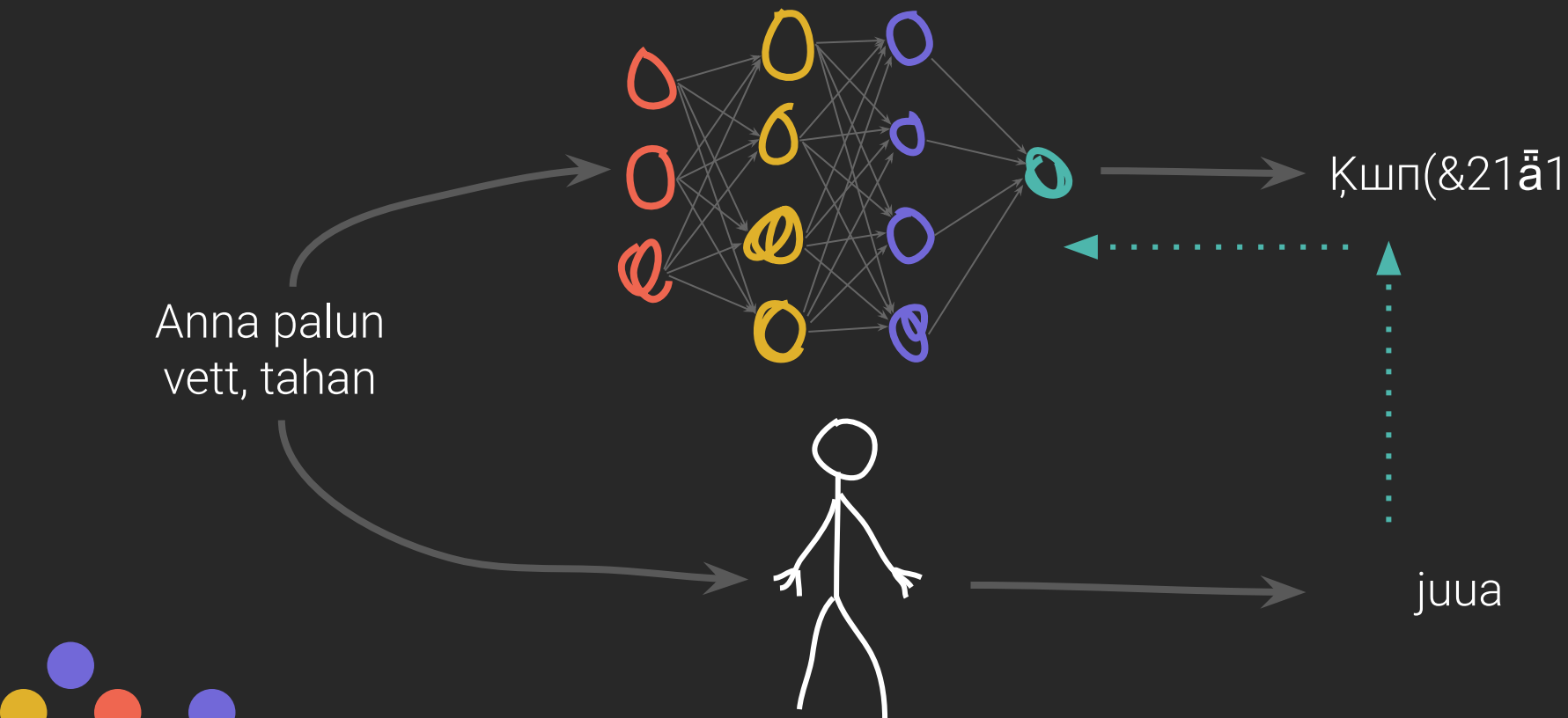
“Anna palun vett, ma tahan juua.” → “⟨lõpp⟩”

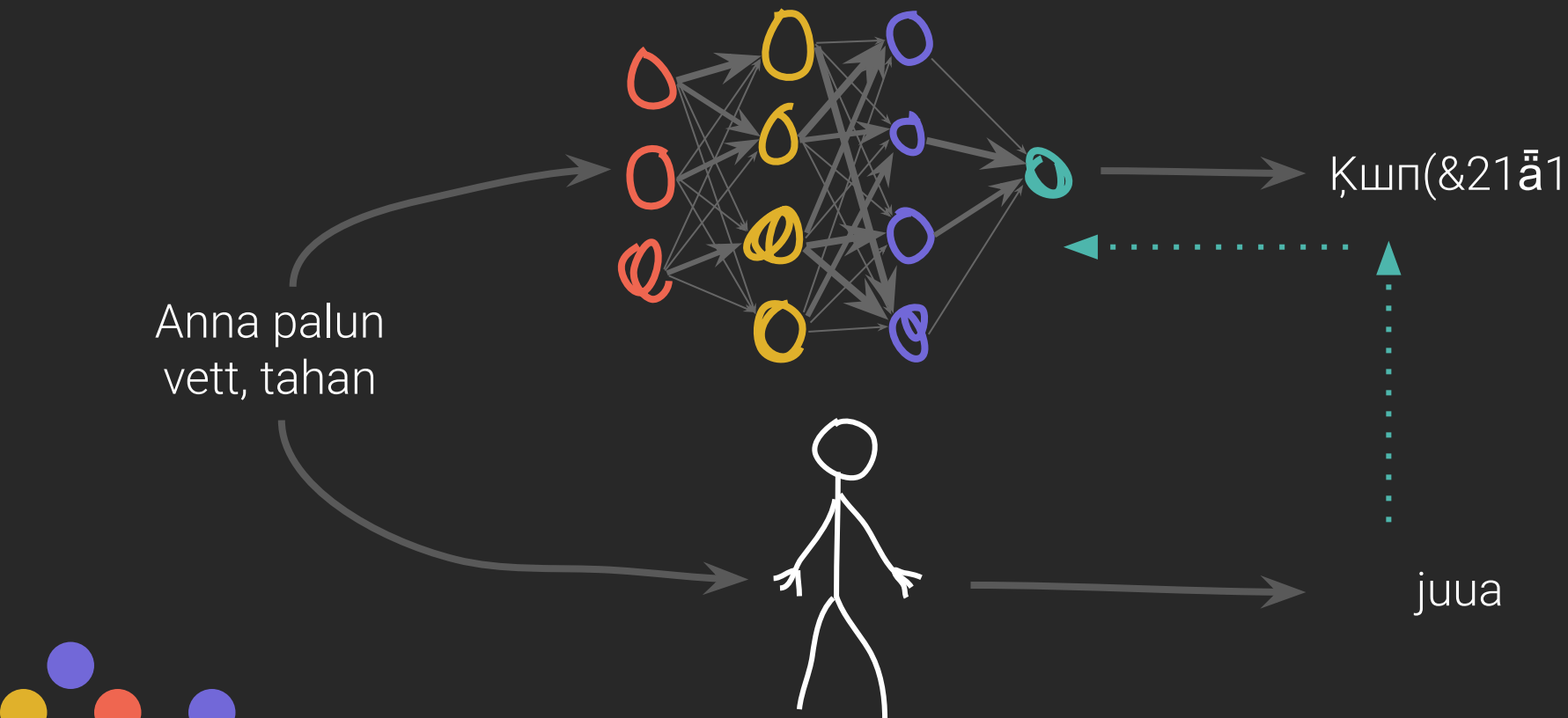


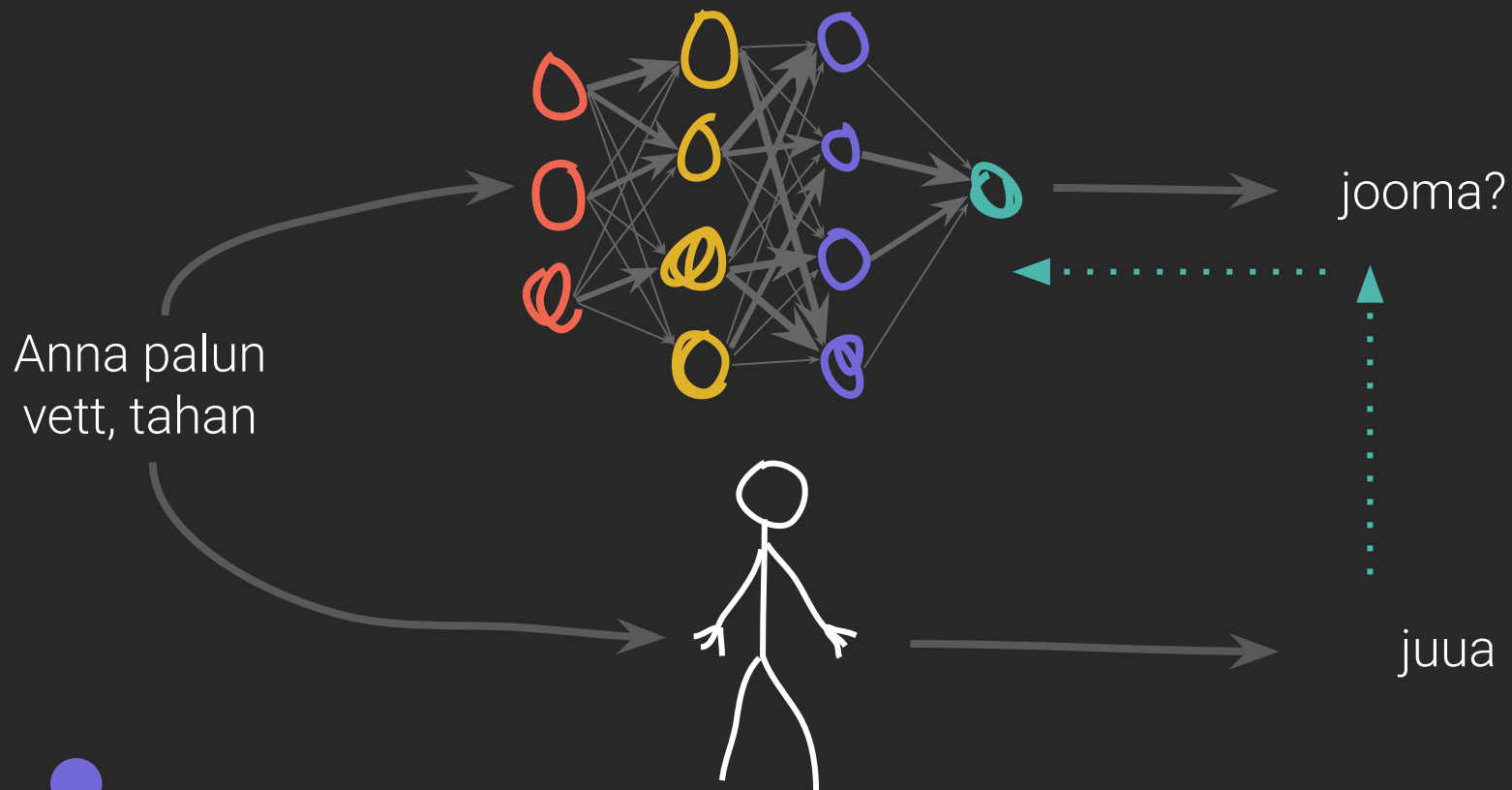


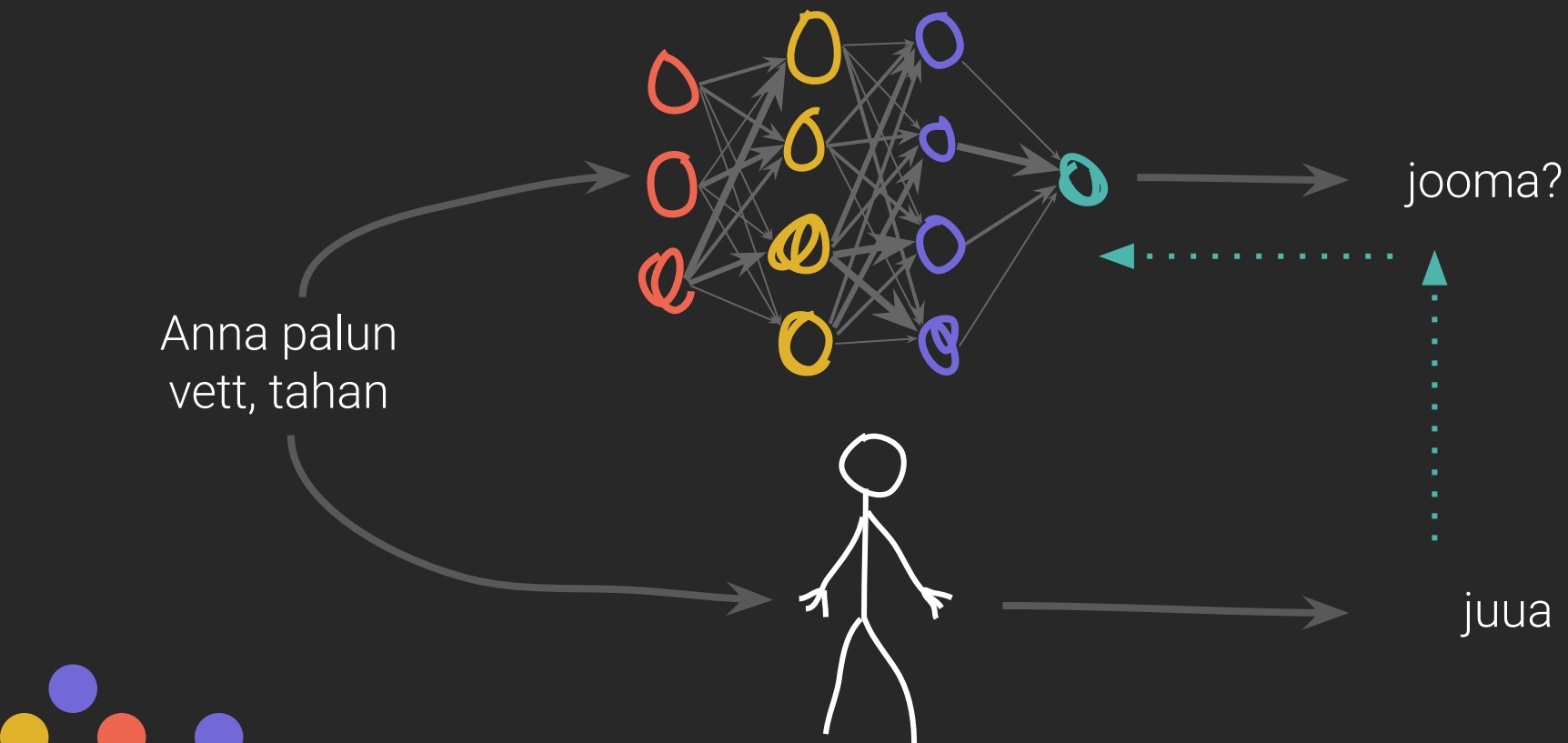


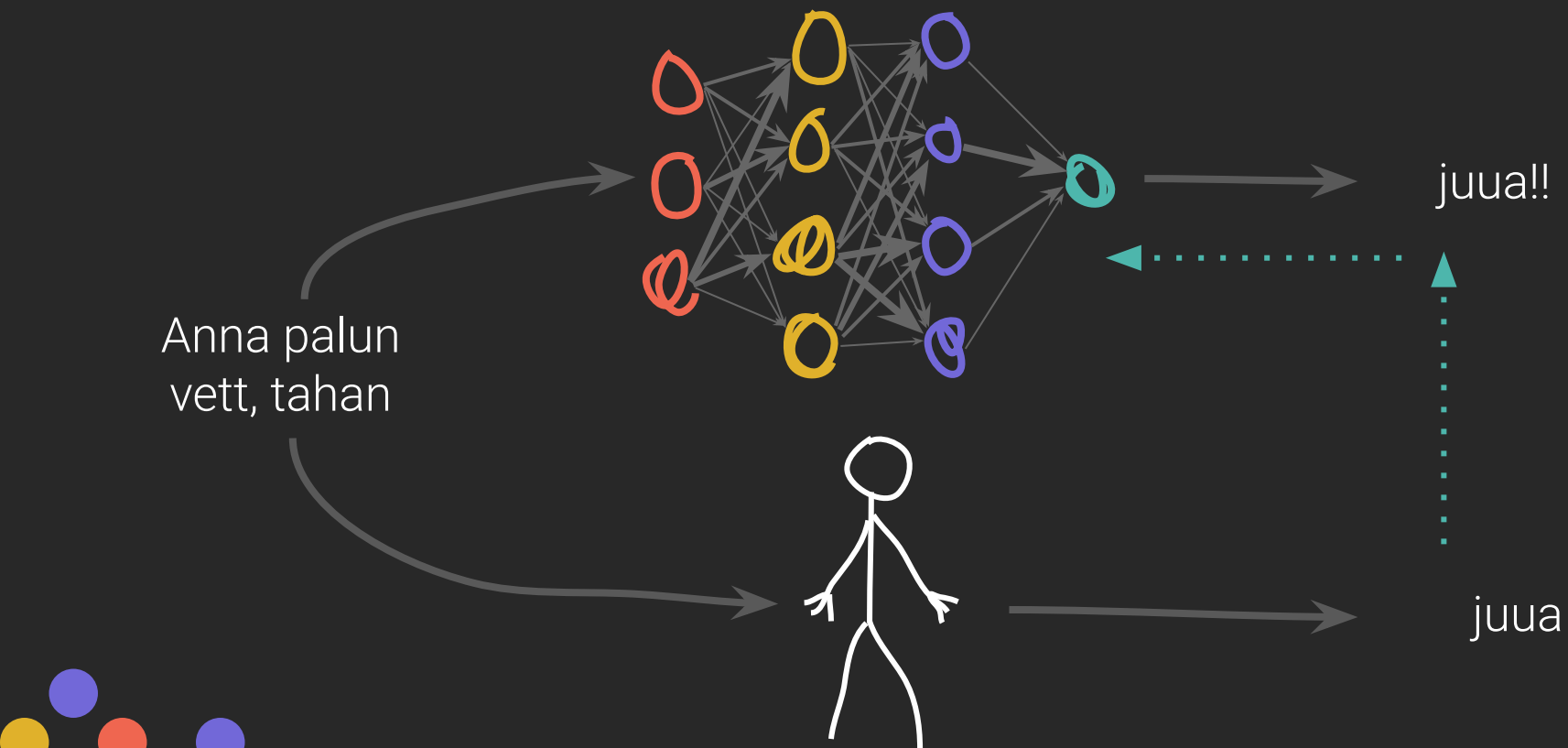












Õppimine

- vajab **andmeid**
 - inglise keel: mitu **triljonit** sõna
 - eesti keel: saadaval on ~4 **miljardit** sõna
 - (“Tõde ja õigus”: ~600 tuhat sõna)
- keele enda eripära (käänded, sõnade järjekord jms) ja erinevus keelte vahel **ei mõjuta eriti tõlke kvaliteeti**
- mõjutab aga **andmete iseloom ja kättesaadavus**
 - eriti tekstivaldkond



Tõlkimine LLMidega

- lihtsalt näitame tõlkepaare, lubame tal õppida
 - kuid siis oskab lõplik mudel **ainult tõlkida**
- me aga tahame samasugust paindlikkust nagu ChatGPT, ehk et saaks kasutada
 - **viibatehnikat** (promptimist)
 - **interaktiivsust**
 - **kohanemist kontekstile**



Prompting / juhised

Pluss: Keelemudeliga masintõlkimist saab **suunata** promptimise abil



ChatGPT ▾



FI



Tõlgi tekst "Naine jõudis koju. Ta jättis oma auto garaaži ning läks tuppa." eesti keelest inglise keelde. Tõlkimisel kasuta vaba teksti stiili ja kõnekeelt.



The woman got home. She parked her car in the garage and went inside.



Message ChatGPT



ChatGPT can make mistakes. Check important info.



chat.openai.com

Interaktiivsus

Pluss: keelemudel saab oma tõlkeväljundit **ise parandada**



Tõlgi lause "I stopped at the bank" eesti keelde.



Lause "I stopped at the bank" tõlgituna eesti keelde on: **"Ma peatusin panga juures."**

Paranda tõlget, arvestades, et jutt käib jõe kohta, ehk river bank. Vastuseks ainult parandatud tõlge



Ma peatusin jõe kaldal.



Message ChatGPT



Mitmekeelsus

Mitmekeelsus ja väiksema rääkijate/resssursside hulgaga keeled
ei ole keelemudelites nii tugevalt toetatud



Keelemudelid

Madalam efektiivsus ja usaldusväärsus võrreldes lause haaval tõlkimisega.



Keelemudelid

Madalam efektiivsus ja usaldusväärsus võrreldes lause haaval tõlkimisega.

(2016: statistilised mudelid on kiiremad, efektiivsemad ja usaldusväärsemad)



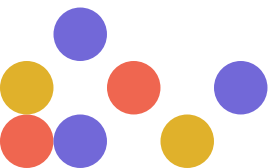
Soomõ-ugri keeled

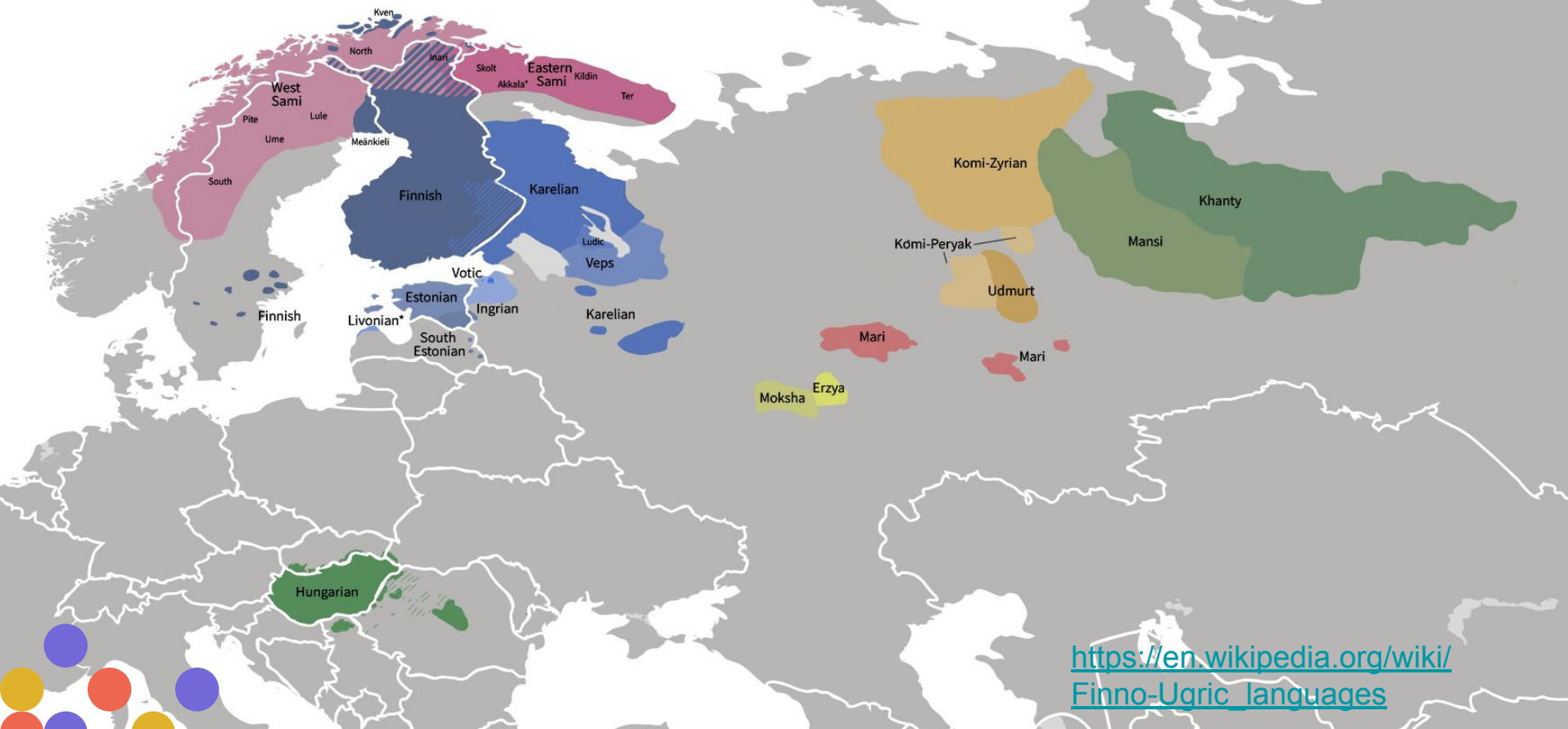
Suopmelaš-ugralaš gielat

SMUGRI: Finno-Ugric NLP

Суоми-Угрань кяльхне

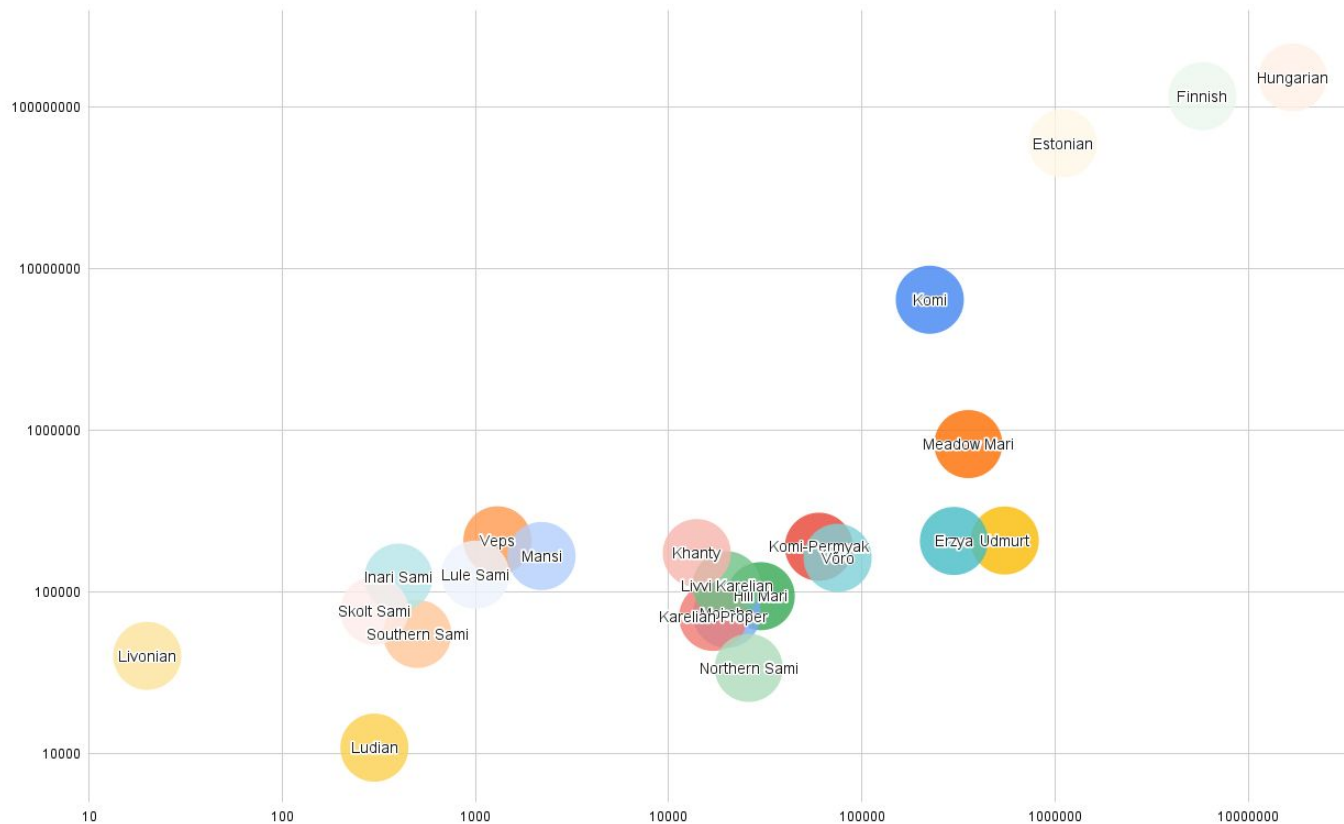
Suomalais-ugrilaiset kielet





https://en.wikipedia.org/wiki/Finno-Ugric_languages

Kõnelejate arv / andmed



Projekti keeled

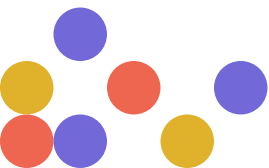
- 2021: võru - põhjasaami - lõunasaami - eesti - soome masintõlge
- 2022: liivi - eesti - läti - inglise masintõlge
+ inarisaami, skoltsaami, lulesaami
- 2023: + vepsa, udmurdi, ersa, handi, mansi, päriskarjala, livviko,
lüüdi, mäe- ja niidumari, mokša, sürja- ja permikomi masintõlge

+ võrukeelne kõnesüntees
- 2024, plaanis: lisada isuri, vadja, pite saami ja kildin saami keeli



Tulemused

- vabavaraline tõlkesüsteem 20+ soome-ugri keelele
- avalik andmestik tõlgete ja niisama tekstidega
- veebidemo: translate.ut.ee



Sisendkeel : inglise



Väljundkeel : liivi

I would like to buy a house



...



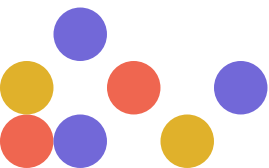
27 / 7000



Võrdle tõlkeid

Väljakutsed

- keel?
 - vastunäited: nt isuri
- ortograafia?
 - vastunäited: nt handi
- andmed? kõnelejad?
 - vastunäited: enamik soome-ugri perest



Lõplikud eesmärgid

Soome-ugri ja Eesti keelte ja murrete

- masintõlge
- keelemudelid
- kõnesüntees ja transkribeerimine
- keeleõppe abiprogrammid



Sõnum kaasa

- masintõlge ja tehisintellekt pidevalt arenevad
- nad ei asenda inimesi, kuid ikka tugevalt mõjutavad meie elu ja tööd
- tasub olla kursis ja pidevalt õppida
- keelemudelid õpivad ja töötavad väga lihtsalt kombel, kuid saavutavad muljetavaldatavaid tulemusi
 - mida aga ei tasuks üle hinnata
 - ega ka alahinnata



Aitäh!

<https://tartunlp.ai>