

EESTI KEELETEHNOLOOGIA 2021

Einar Meister

TTÜ keeletehnoloogia labor

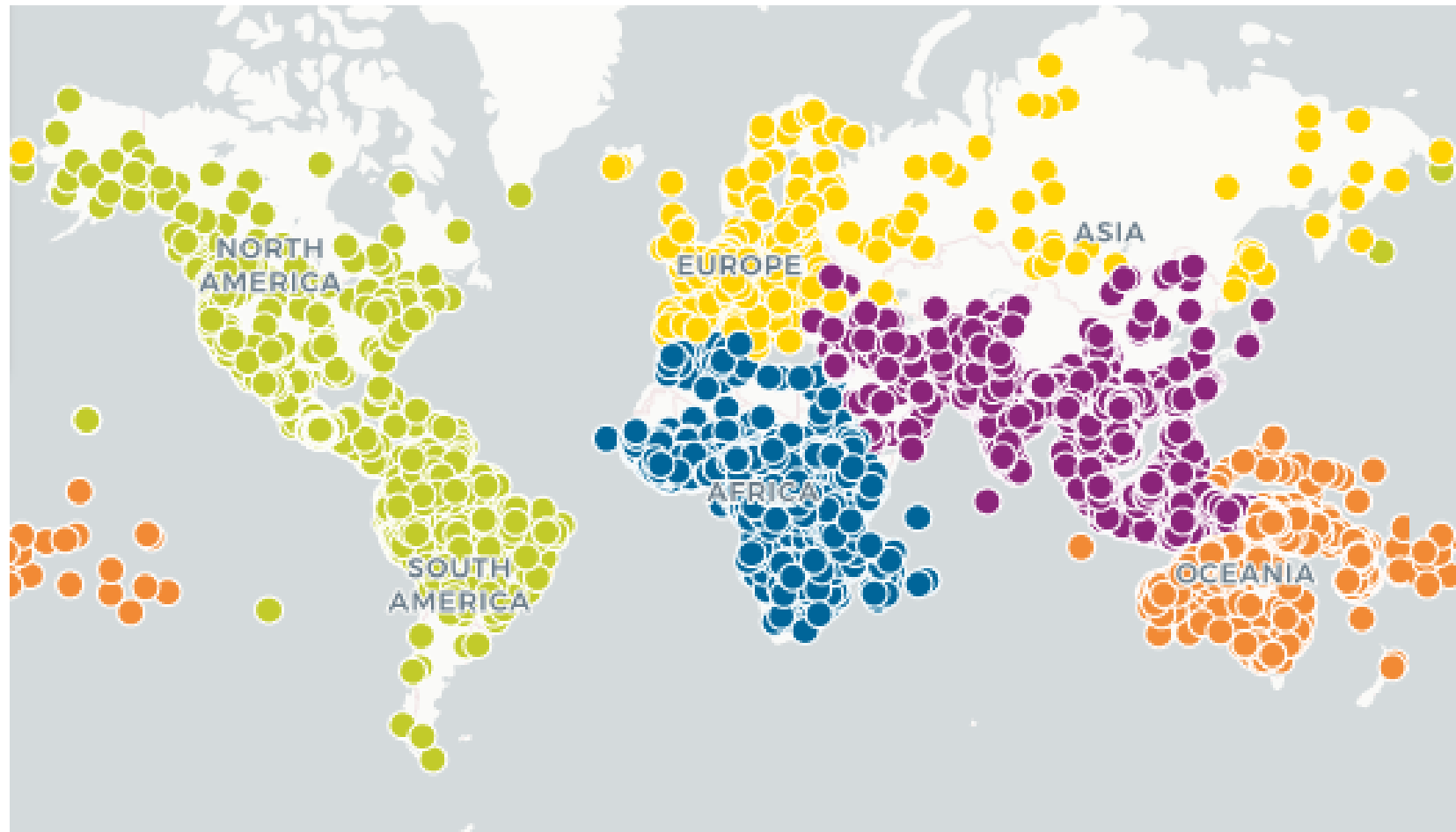
einar.meister@ttu.ee

Ettekande sisu

- Keelete ja tehnoloogia areng
- Mis on keeletehnoloogia?
- Keeletehnoloogia arendus Eestis
- Tehnoloogilised lahendused
- Võrdlus teiste keeltega
- Kokkuvõte

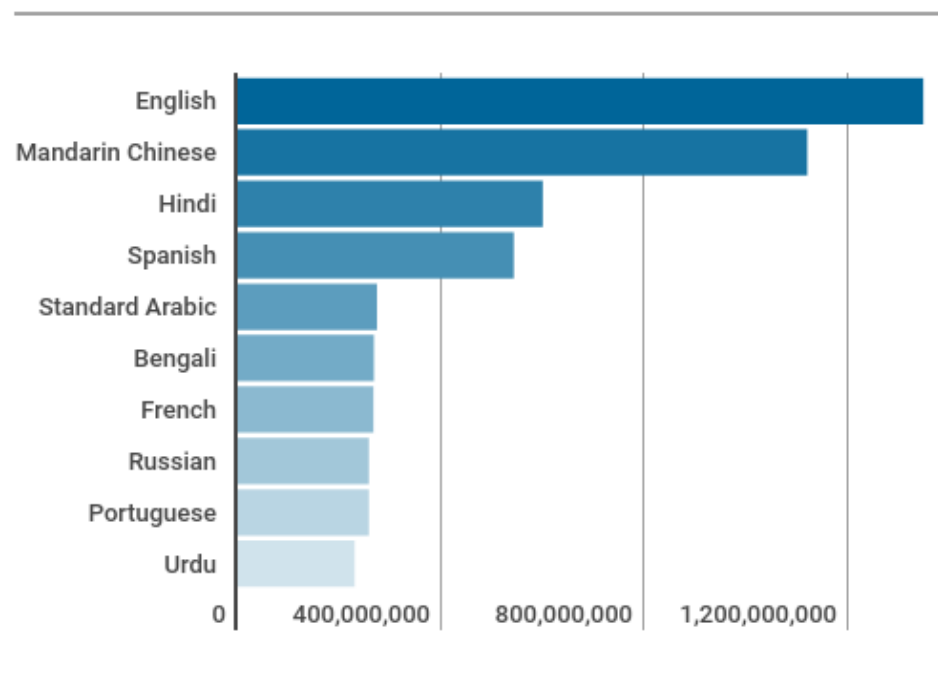
Maailma keeled

<https://www.ethnologue.com/>



10 suuremat keelt

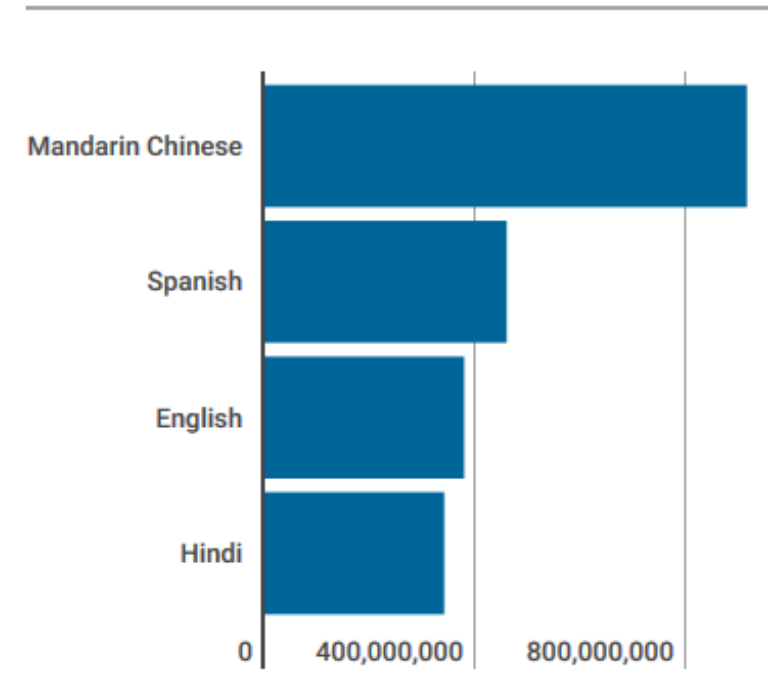
Top 10 most spoken languages, 2021



 Ethnologue

<https://www.ethnologue.com/guides/most-spoken-languages>

Languages with the most native speakers



Suured keeled, väikesed keeled

Table 2. Distribution of world languages by number of first-language speakers

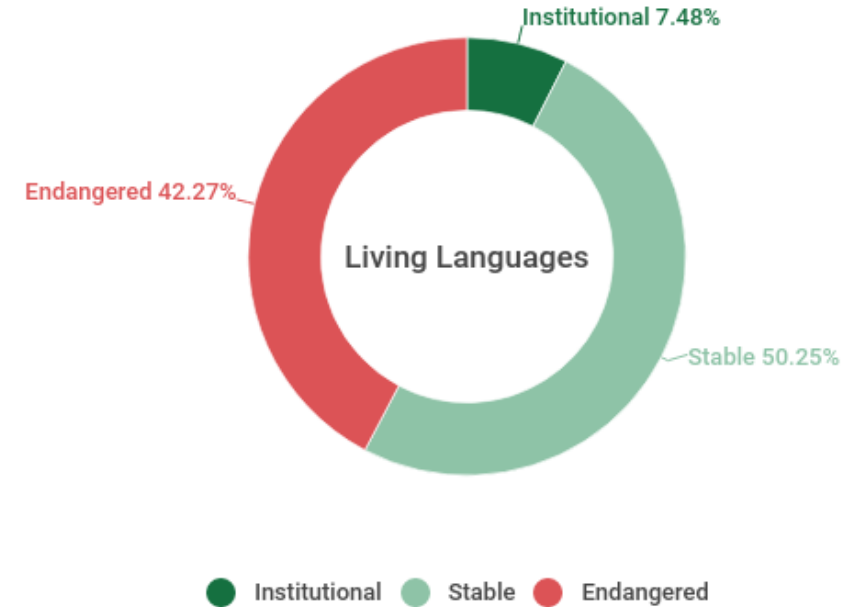
Population range	Living languages			Number of speakers		
	<i>Count</i>	<i>Percent</i>	<i>Cumulative</i>	<i>Total</i>	<i>Percent</i>	<i>Cumulative</i>
100,000,000 to 999,999,999	8	0.1	0.1%	2,614,031,200	40.17718	40.17718%
10,000,000 to 99,999,999	84	1.2	1.3%	2,608,411,777	40.09081	80.26798%
1,000,000 to 9,999,999	306	4.3	5.6%	917,412,798	14.10047	94.36845%
100,000 to 999,999	944	13.3	18.9%	297,092,808	4.56626	98.93471%
10,000 to 99,999	1,808	25.5	44.4%	61,213,595	0.94084	99.87555%
1,000 to 9,999	1,979	27.9	72.3%	7,614,348	0.11703	99.99258%
100 to 999	1,070	15.1	87.3%	469,345	0.00721	99.99980%
10 to 99	337	4.7	92.1%	12,753	0.00020	99.99999%
1 to 9	132	1.9	94.0%	536	0.00001	100.00000%
0	220	3.1	97.1%	0	0.00000	100.00000%
Unknown	209	2.9	100.0%			
<i>Totals</i>	7,097	100.0		6,506,259,160	100.00000	

Suured keeled, väikesed keeled

- 200 suurema keele kõnelejad moodustavad ca 88% maailma rahvastikust
- Üle 1 miljoni kõnelejaga keeled (ca 400 keelt) moodustavad 5,6% kõigist maailma keeltest; nende kõnelejate osakaal on 94% maailma rahvastikust
- Kas eesti keel on suur või väike?
- Suured keeled: > 1 miljon kõnelejat
- Keskmise suurusega keeled: 10 tuhat – 1 miljon kõnelejat
- Väikesed keeled: < 10 tuhat kõnelejat

Ohustatud keeled

- Igal aastal sureb välja ca 40 – 50 keelt
- Ohustatud keelete arv:
 - 2019: 2895
 - 2021: 3018



Ohustatud keeled: 3018



Keelte tulevik

- Prognoos: 100 aasta pärast on maailmas kasutusel vaid 50-10% täna eksisteerivatest keeltest
- Gutenbergi efekt: trükikunst suretas välja keeled, mille puhul ei olnud olemas kirjakeelt
- IT mõju: keeli, millel puudub arvutitugi, ootab ees digitaalne hääbumine
- Kultuurkeeled – on olemas emakeelne piibel (~2400 keelt)
- Tehnoloogiliselt jätkusuutlikud keeled – on olemas keele masintöötamise tehnoloogia (~ 1-2%)

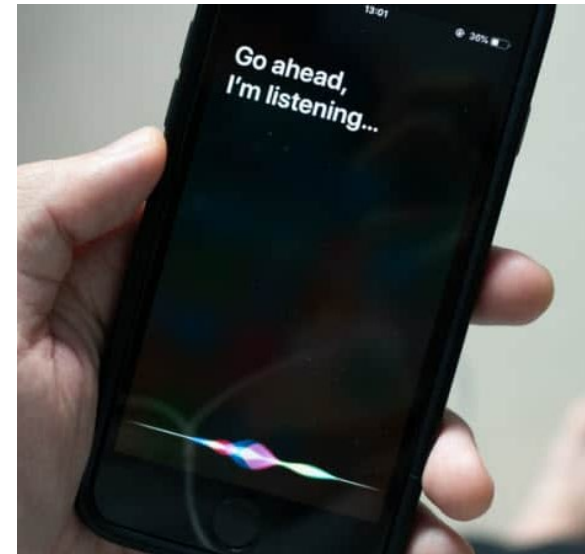
Infotehnoloogia areng

Thomas J. Watson (IBM president), 1943: „Ma arvan, et maailmas on turgu vahest ehk viie arvuti jaoks.”

Ken Olson, (Digital Equipment Corp. president), 1977:
„Ma ei näe mingit põhjust, miks inimene peaks endale koju arvutit tahtma.”



Infotehnoloogia areng: liidesed



Virtuaalsed assistendid

2011: Apple Siri

2014: Microsoft Cortana

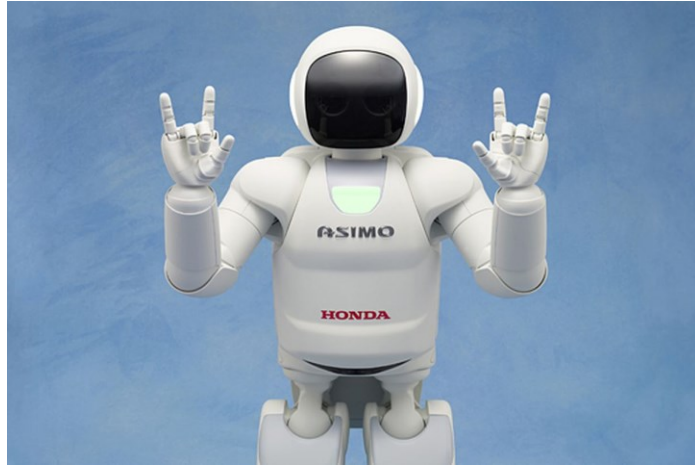
2016: Google Assistant

2014: Amazon Alexa



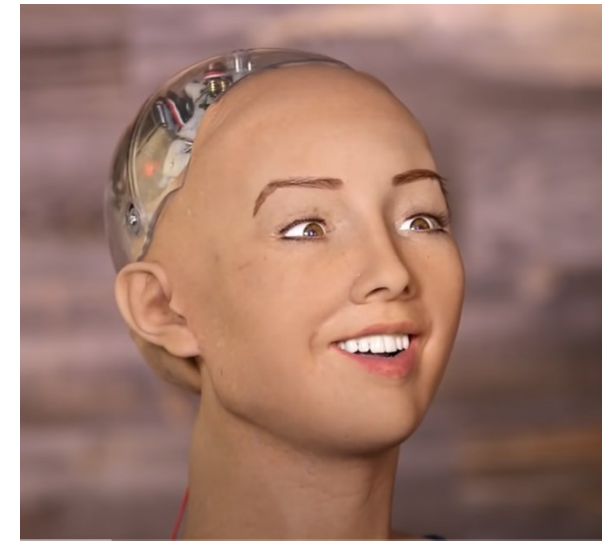
Humanoidrobotid

Asimo

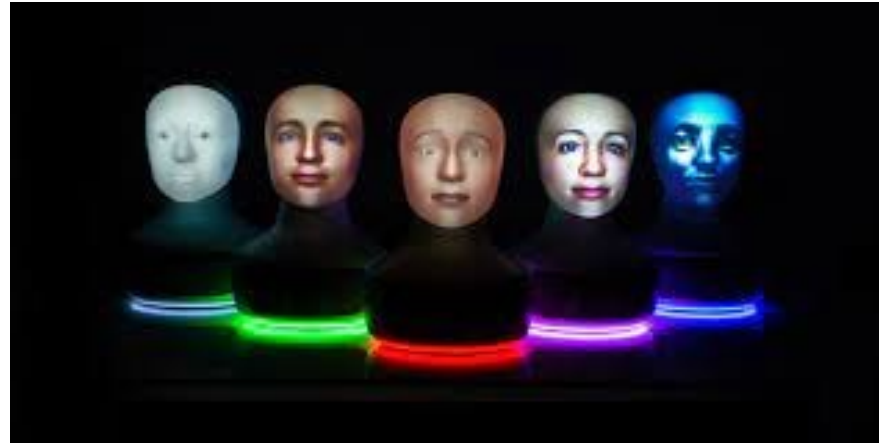


Furhat <https://www.youtube.com/watch?v=WlQc0Tbzbms>
<https://www.youtube.com/watch?v=k5Zc8lO8hec>

Sophia https://www.youtube.com/watch?v=W0_DPi0PmF0
<https://www.youtube.com/watch?v=LguXfHKsa0c>



Algamas on inimene-masin ajastu



THE DAWN OF THE HUMAN-MACHINE ERA

A FORECAST OF NEW AND EMERGING LANGUAGE TECHNOLOGIES

<https://lithme.eu/wp-content/uploads/2021/05/The-dawn-of-the-human-machine-era-a-forecast-report-2021-final.pdf>

Keeletehnoloogia on keelespetsiifiline

- Ingliskeelne kõnesüntees ei sobi eesti keele jaoks
“Üle oja mäele, läbi oru jõele.”



- KT-suutlikud keeled on eelkõige suure kõnelejate arvuga keeled (> 50 milj. kõnelejat) – keeletehnoloogia arendamine on kallis!
- Keelte puhul, mille kõnelejate arv on alla 10 miljoni, on keeletehnoloogia arendustöö majanduslikult ebaotstarbekas
- Aga 1 miljoni kõnelejaga keeled?

Mitmekeelsus Euroopa Liidus

- EL on olemuslikult mitmekeelne – kõik keeled on võrdsed
- Digitaalne lõhe:
 - suured keeled – majanduslikult huvipakkuvad keeled
 - väikesed keeled:
 - väike kõnelejate arv
 - piiratud tehnoloogiline tugi
 - majanduslikult perspektiivitud
- Puudub spetsiifiline programm mitmekeelsuse tehnoloogiliseks arenduseks
- Ei jätku ressursse kõigi EL ametlike keelte tehnoloogilise toe arendamiseks
- **Kus on väljapääs?**
 - Ainult üks ametlik keel – inglise keel
 - Rahvuslikud/riiklikud programmid + rahvusvaheline koostöö

Euroopa Parlamendi resolutsioon (11.09.2018)

Keeleline võrdõiguslikkus digitaalajastul

Praegused takistused keelelise võrdõiguslikkuse saavutamisel digitaalajastul

Euroopa Parlament ...

- peab kahetsusväärseks, et asjakohase poliitika puudumise tõttu **suureneb Euroopas praegu tehnoloogialõhe hästi rahastatud ja vähem rahastatud keelte vahel**
- avaldab kahetsust, et **rohkem kui 20 Euroopa keelt on digitaalses väljasuremisohus**
- juhib tähelepanu sellele, et **digilõhele laialdaselt kasutatavate ja vähemkasutatavate keelte vahel suureneb**
- märgib, et **keeletehnoloogia on kõigepealt kättesaadav inglise keeles**
- märgib, et **keeletehnoloogial ei ole Euroopa poliitilises tegevuskavas tähtsust**, hoolimata asjaolust, et keelelise mitmekesisuse austamine on sätestatud aluslepingutes
- tunnustab selliste eelnevate ELi rahastatud teadusuuringute võrgustike nagu **FLaReNet, CLARIN, HBP** ja **META-NET** (sealhulgas META-SHARE) rolli, mis sillutasid teed Euroopa keeletehnoloogia platvormi väljaarendamiseks

Euroopa Parlamendi resolutsioon (11.09.2018)

Keeleline võrdõiguslikkus digitaalajastul

Keeletehnoloogiapoliitika institutsioonilise raamistiku parandamine ELi tasandil

- nõuab, et nõukogu koostaks soovitusel liidu kultuurilise ja keelelise mitmekesisuse kaitsmise ja edendamise kohta – ka keeletehnoloogia valdkonnas;
- julgustab neid liikmesriike, kus on juba välja töötatud edukad keeletehnoloogia valdkonna poliitikastrateegiad, oma kogemusi ja häid tavasid jagama
- palub komisjonil ja liikmesriikidel määrata minimaalsed keeleressursid, mis peaksid kõikidel Euroopa kehtel olemas olema, näiteks andmekogud, leksikonid, kõneandmed, tõlkemälud, annoteeritud korpused ja entsüklopeediline teave, et vältida digitaalset väljasuremist
- kutsub komisjoni üles seadma keeletehnoloogia valdkonnas esikohale need liikmesriigid, mis on väikesed ja millel on oma keel, et aidata lahendada nende keeleprobleeme
- rõhutab vajadust kahandada keeltevahelist tehnoloogialõhet, tugevdades teadmus-ja tehnosiiret

Euroopa Parlamendi resolutsioon (11.09.2018)

Keeleline võrdõiguslikkus digitaalajastul

Soovitud ELi teaduspoliitika kohta

- kutsub komisjoni üles looma ulatuslikku ja pikaajalist rahastamisprogrammi teadus-ja arendustegevuse ning innovatsiooni jaoks keeletehnoloogia valdkonnas Euroopa, riiklikul ja piirkondlikul tasandil
- teeb ettepaneku luua koos kõikide Euroopa keelte esindajatega Euroopa keeletehnoloogia platvorm, mis võimaldaks jagada keeletehnoloogia ressursse, teenuseid ja avatud lähtekoodiga pakette
- soovitab luua või laiendada selliseid projekte nagu digitaalse keelte mitmekesisuse projekt (**Digital Language Diversity Project**), mille raames tehakse uurimistööd kõigi Euroopa keelte digitaalsete vajaduste kohta
- teeb ettepaneku ajakohastada META-NETi valge raamatu seeriat, mis avaldati 2012. aastal üleeuroopalise keeletehnoloogia uuringu vormis, võimaldamaks keeletehnoloogia poliitikat hinnata ja arendada

EL projektid

- **META-NET** – keeletehnoloogia võrgustik (60 partnerit 34st riigist) <http://www.meta-net.eu/>
- **CLARIN** – digitaalne infrastruktuur keeleressurssidel põhineva uurimistöö toetamiseks <https://www.clarin.eu/>
- **European Language Resource Coordination** – haldab ja koordineerib keeleressursse kõigis ELi ja Euroopa Ühendamise Rahastu assotsieerunud riikide ametlikes keeltes <https://www.lr-coordination.eu/>
- **European Language Grid** – keeletehnoloogia platvorm kõigi Euroopa keelte jaoks (rakendused, teenused, andmekogud) <https://www.european-language-grid.eu/>

Eesti keel ja tehnoloogia

- Riiklik programm “Eesti keele keeletehnoloogiline tugi (2006-2010)”
- Riiklik programm “Eesti keeletehnoloogia 2011–2017”
- Riiklik programm “Eesti keeletehnoloogia 2018–2027”
- <https://www.hm.ee/et/tegevused/teadus/teadusprogrammid>
- <https://www.keeletehnoloogia.ee/>
- Keeletehnoloogia programmide eesmärk: luua KT arenduseks vajalikud keeleressursid ja keespetsiifilised tehnoloogilised lahendused

KT uurimine ja arendus Eestis

- Uurimisgrupid:

- Tartu ülikool, arvutiteaduse instituut
<https://nlp.cs.ut.ee/>
- Tallinna tehnikaülikool, keeletehnoloogia labor
<https://taltech.ee/en/laboratory-language-technology>
- Eesti Keele Instituut
<http://portaal.eki.ee/tegevusvaldkonnad.html>
 - Kõneuurimise ja kõnetehnoloogia osakond
 - Keeletehnoloogia kompetentsikeskus

- Eesti Keeleressursside Keskus
<https://keeleressursid.ee/>

- Firmad:

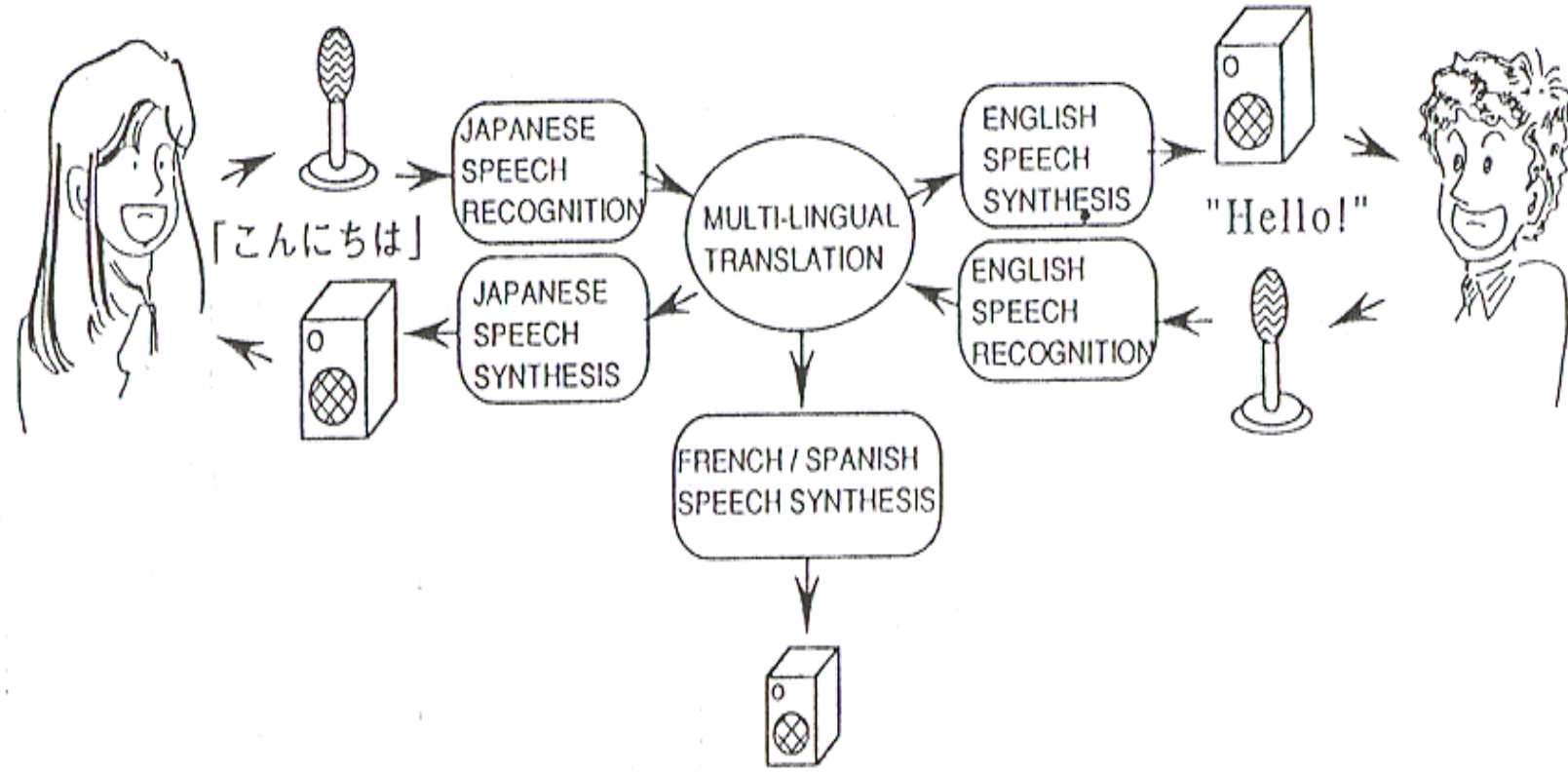
- FiloSoft <http://filosoft.ee/>
- Tilde Eesti <https://tilde.ee/>
- Tarkvara Tehnoloogia Arenduskeskus
<https://www.stacc.ee/et/>
- TEXTA <https://www.texta.ee/>
- Feelingstream OÜ
<https://www.feelingstream.com/>

Keele- ja kõnetehnoloogia

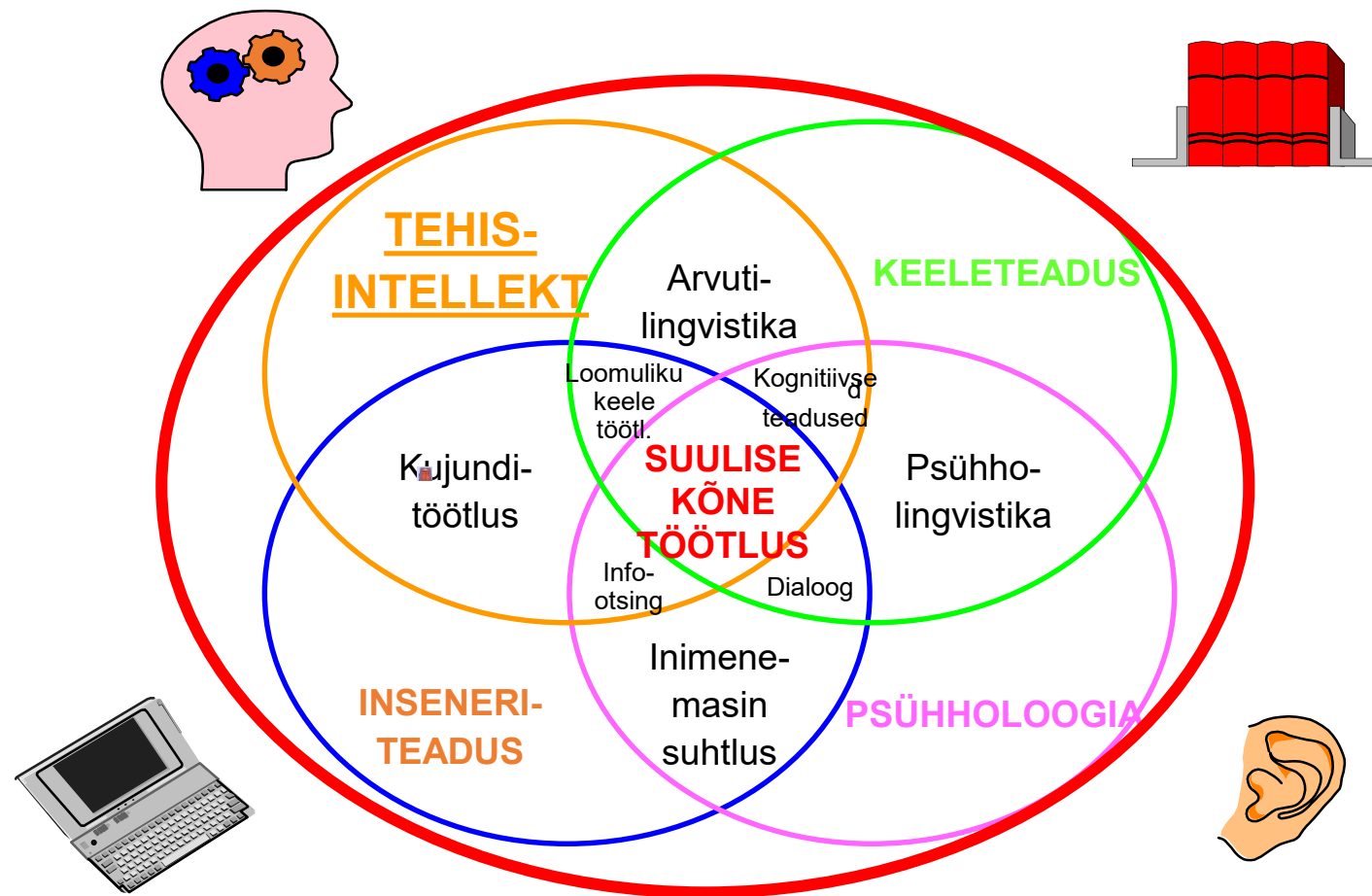
Keeletehnoloogia on keelealaste teadmiste rakendamine arvutisüsteemide loomiseks, mis võimaldavad analüüsida, tuvastada, mõista ja sünteesida inimkeelt kõigis tema vormides

Kõnetehnoloogia – suulise keele töötlus: analüüs, kodeerimine, süntees, tuvastus, jne

Keeletehnoloogia olemus



Keeletehnoloogia on interdistsiplinaarne



Keeletehnoloogia komponendid

Tehnoloogilised lahendused:

kõne süntees ja tuvastus,
morfoloogiline, süntaktiline ja
semantiline analüüs, masintõlge,
keeleõppevahendid, jne

Keeleressursid:

kõne- ja tekstikorpused, elektroonsed
sõnastikud ja andmebaasid, ressursside
loomise ja haldamise vahendid

Teadmised keele ehitusest:

foneetika, fonoloogia, morfoloogia, süntaks,
semantika, pragmaatika

Tehnoloogilised lahendused (1)

- Morfoloogia – sõnavormide analüüs ja süntees

mees

mees+0 //_S_ sg n, //

mesi+s //_S_ sg in, //

Rakendused: õigekirja kontroll, kõnesüntees ja –tuvastus

- Süntaktiline e grammatiline analüüs – sõnavormide süntaktilise funktsiooni määramine lauses

Ma õppima eesti keel -> Ma õpin eesti keelt

Rakendused: grammatika kontroll, infootsing, tõlkesüsteemid, kõnesüntees ja –tuvastus, automaatne sisukokkuvõtete tegemine

Tehnoloogilised lahendused (2)

- Semantiline analüüs – semantiliste seoste leidmine tekstis

*traktori nahast kate -> (traktori
(nahast kate))*

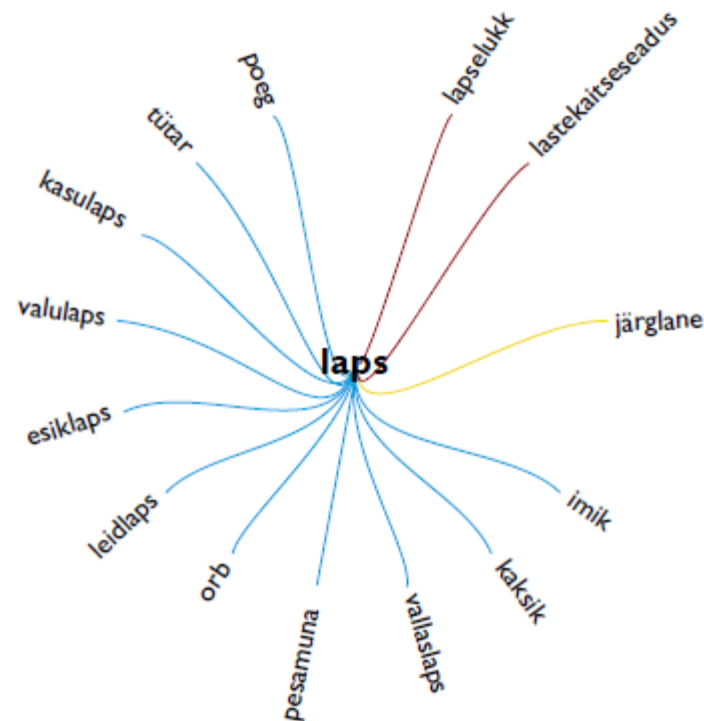
*professori nahast portfell -> (professori
(nahast portfell))*

*sea nahast portfell -> ((sea nahast)
portfell)*

Rakendused: infootsing, semantiline web,
inimene-masin dialoog

Eesti keele teaurus

<http://www.cl.ut.ee/ressursid/teksaurus/>



Tehnoloogilised lahendused (3)

- Pragmaatika – tegeleb suhtluse seadusparasustega, mis on olulised näiteks inimese-masina dialoogi modelleerimisel

Suhtlusaktid dialoogis: küsimused, vastused, selgitused jm

Inimene: Mis kell väljub järgmine rong Narvast Tallinnasse?

Dialoogsüsteem peab:

1. teisendama kõne tekstiks
2. leidma tekstist:
 - mida tahetakse teada (kellaaeg)
 - tegevus (väljumine)
 - sõiduvahend (rong)
 - lähtejaam (Narva)
 - sihtjaama (Tallinn)
3. sooritama päringu Elroni sõiduplaanis
4. formuleerima vastuse

Masin: *Ei saanud aru, palun korda küsimust*

Inimene: *Tõmba uttu!*

Tehnoloogilised lahendused (4)

- Masintõlge – automaatne tõlge eesti keelest ja eesti keelde

<http://translate.google.com/>

<https://neurotolge.ee/>

<https://translate.tilde.com/>

Rakendused: piiritletud valdkonnas töötav tõlkesüsteem, mitmekeelne infootsing

Tehnoloogilised lahendused (5)

- Kõnesüntees – ortograafilise teksti teisendamine inimkõneks
- Sünteesidemo:
 - <https://www.eki.ee/heli/>
 - <https://neurokone.ee/>

Rakendused: <https://www.eki.ee/heli/index.php/rakendused>

- Veebisõnastike helindamisliides
- Nutitelefoni rakendused
- Heliraamatud
- [Subtiitrite helindamine](#)

Tehnoloogilised lahendused (6)

- Kõnetuvastus – kõne teisendamine kirjalikuks tekstiks
- Rakendused:
 - radioloogia (koostöös Põhja-Eesti Regionaalhaiglaga)
 - meediamonitooring
 - Riigikogu istungite ja koosolekute transkribeerimissüsteem
 - avalikud veebiteenused <https://bark.phon.ioc.ee/>:
 - kõnesalvestuste brauser
 - veebipõhine kõnetuvastus
 - dikteerimine reaalajas
 - Androidi-rakendused
 - lähiajal tulemas ERRi saadete automaatsed subtiitrid

Kõnetuvastus

- Kõnetuvastuse veebiteenused: kõnesalvestuste brauser
<http://bark.phon.ioc.ee/tsab/p/index>
 - Automaatse kõnetuvastuse abil transkribeeritud raadiosaadete salvestused alates 2011
 - Suudab hääle järgi identifitseerida u 1800 Eesti avaliku elu tegelast

Kõnetuvastus

- Kõnetuvastuse veebiteenused: veebipõhine kõnetuvastus
<http://bark.phon.ioc.ee/webtrans>
- <http://tekstiks.ee>
- Nädalas transkribeeritakse ca 400 helifaili
- Ca 2000 kasutajat
- Kasutajad:
 - ajakirjanikud (intervjuude transkribeerimine)
 - humanitaarteadlased, üliõpilased
 - pressiosakonnad (nt valitsuse pressikonverentside transkribeerimine)

Kõnetuvastus

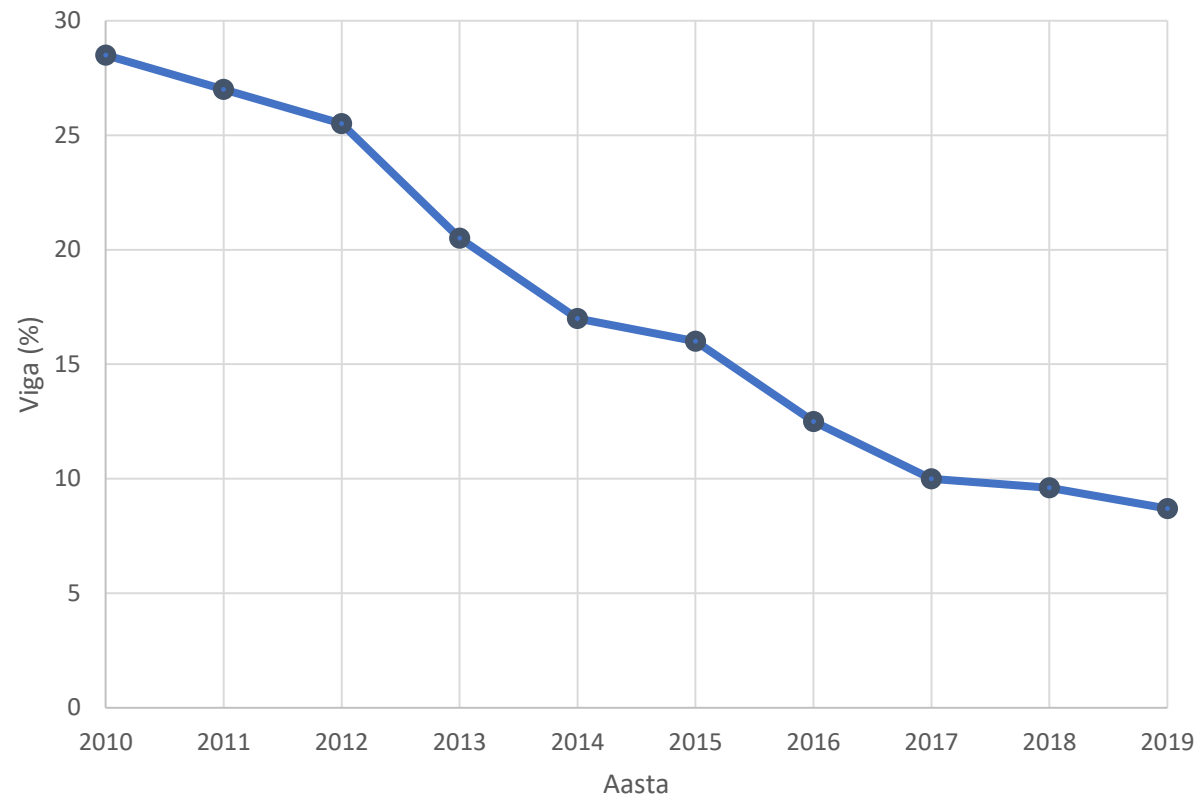
- <http://tekstiks.ee>
- Kõnesalvestuste transkribeerimise ja transkriptsioonide haldamise veebiteenus
- Funktsionaalsus:
 - Salvestuste üleslaadmine ja automaatselt tuvastatud transkriptsioonide vaatamine/kuulamine koos sünkroniseeritud heliga
 - Teksti järeltoimetamine
 - Teksti eksport

Kõnetuvastus

- Kõnetuvastuse veebiteenused: kõnetuvastus reaalsajas
<https://bark.phon.ioc.ee/dikteeri>

Kõnetuvastus

Kõnetuvastuse viga protsent:



Kõnetuvastus

Kõnetuvastus $\neq$ kõne mõistmine

Suuline dialoog arvutiga eeldab mitmete keeletesiifiliste komponentide olemasolu – kõnetuvastus, sõnumi mõistmine, vastuse genereerimine, kõnesüntees

Kõnetuvastuse probleemid:

- töötab hästi piiratud sõnavaraga valdkonnas, nt radioloogia
- töötab halvasti (1) spontaanse kõne, (2) võõrkeelse aktsendiga ja koodivahetusega kõne, (3) laste kõne puhul
- ei tööta üldse mitme inimese samaaegse kõne puhul

Tehnoloogilised lahendused (7)

- Kõnetõlge – kõne teisendamine kirjalikuks tekstiks ja tõlkimine teise keelde
- <https://live.neurotolge.ee>

Tehnoloogilised lahendused (8)

- Kõnelejatuvastus:
 - identifitseerimine
 - verifitseerimine
- Kõneldava keele tuvastus – mis keeles kõneldakse
- Abivahendid kõnepuuetega inimestele
- Paralingvistilised rakendused:
 - kõneleja soo ja vanuse identifitseerimine
 - stressi, depressiooni jm haiguste tuvastamine
 - emotsioonide tuvastus
- Keeleõppe rakendused:
 - võõrkeeke hääldustreening <https://calst.hf.ntnu.no/>
 - eesti keele hääldustreening – mõned prototüübid TTÜ lõputööde raames
 - lugemise karaoke <https://readingkaraoke.xyz/>

Korpused

- Kaasaegne keeletehnoloogia põhineb valdavalt masinõppel – treenimiseks on vajalikud suuremahulised märgendatud korpused
- Korpus on eesmärgipäraselt koostatud elektroonilise (arvutis töödeldava) keeleainese kogum, mis kirjeldab võimalikult tõepäraselt tegelikku keelekasutust
- Tekstikorpused sisaldavad suurt hulka eriliigilisi tekste
- Kõnekorpused sisaldavad paljude inimeste kõnenäiteid

Eesti Keeleressursside Keskus (EKRK)

- EKRK (<https://keeleressursid.ee/>) arhiveerib ja teeb kättesaadavaks eesti keele digitaalsed ressursid ja tehnoloogiad (keeletarkvara, sõnastikud, teksti- ja kõnekorpused, keeleandmebaasid)
- Keeleressursside otsing <https://metashare.ut.ee/>
- Korpuste päringusüsteem [KORP](https://korp.keeleressursid.ee/) – otsing tekstikorpustes <https://korp.keeleressursid.ee/>
- Ühendatud sisuotsing [RABA](https://raba.keeleressursid.ee/) – tekstikorpused, kõnekorpused ja leksikonid <https://raba.keeleressursid.ee/>

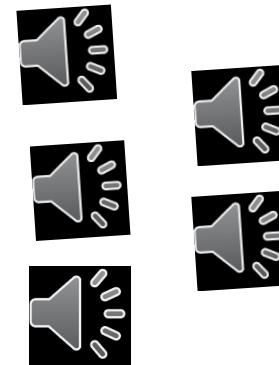
TTÜ keeletehnoloogia labori korpused

- BABEL-Eesti
- SpeechDat-Eesti
- Raadiouudiste korpus
- Intervjuude korpus
- Loengukõne korpus
- Aktsendikorpus
- Lastekõne korpus
- Seenioride korpus
- Koosolekute korpus
- ...

TTÜ keeletehnoloogia labori korpused

Aktsendikorpus

- 18 erineva emakeelega eesti keele kõnelejate salvestused
 - vene 50
 - soome 30
 - läti 20
 - saksa 15
 - prantsuse 13
 - leedu 12
 - jaapani 7
 - ...
 - 20 eesti emakeelega kõnelejat
- Vanus: 18-62 aastat
- Alustasid eesti keele õppimist 16-47 aastaselt
- Ca 15 min kõnet igalt kõnelejalt



Eesti KT võrdlus

- META-NET: Euroopa põhiliste keelte tehnoloogilise toe võrdlus (2012)
<http://www.meta-net.eu/whitepapers/overview>
 - Masintõlge
 - Kõnetöötlus
 - Tekstianalüüs
 - Kõne- ja tekstikorpused
- Võrdluses 30 keelt

Tehnoloogilise toe pingerida 2012

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian (Bokmål,Nynorsk) Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese Welsh

<http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison>

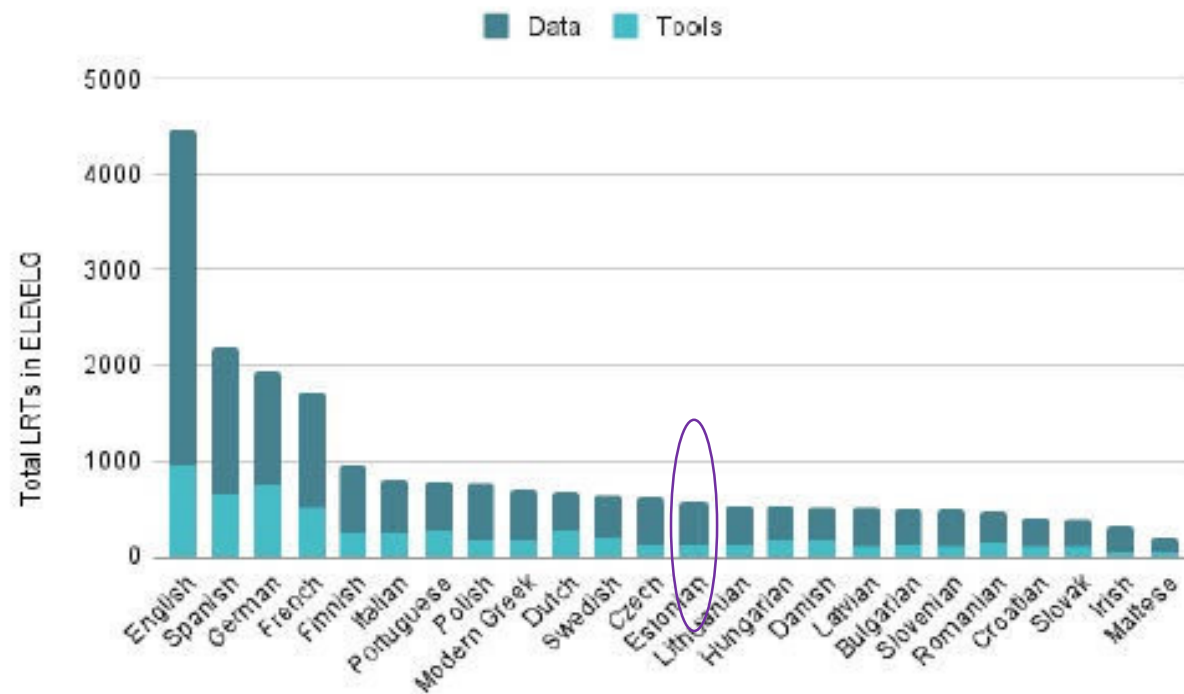
Tehnoloogilise toe pingerida 2012



Käimasolev analüüs

- META-FORUM 15.-17.11.2021
<https://www.european-language-grid.eu/meta-forum-2021/>
- Euroopa keelte digitaalse võrdõiguslikkuse analüüs (tulemused 2022)

Esmased tulemused:



Kokkuvõtteks

- Eesti keele tehnoloogiline tugi on maailmatasemel, vähemalt 50 esimese hulgas!
- Riigi tugi keeletehnoloogia programmi näol on taganud stabiilse arengu
- On loodud suur hulk vajalikke keeleressursse ja prototüüpe
- Mitmed edukad rakendused avalikus ja erasektoris, IT-firmadel on tekkinud huvi keeletehnoloogia kasutamise vastu
- Rohkem ressursse on vaja uurijate-arendajate järelkasvu koolitamiseks
- **Eesti keele digitaalset hääbumist pole lähikümnetel põhjust karta!**