EESTI KEELETEHNOLOOGIA ANNO 2013

Einar Meister
TTÜ Küberneetika Instituut
Foneetika ja kõnetehnoloogia labor

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

Infotehnoloogia areng: liidesed

Microsoft (1998): "Klaviatuurita arvuti tuleb viie aasta pärast"

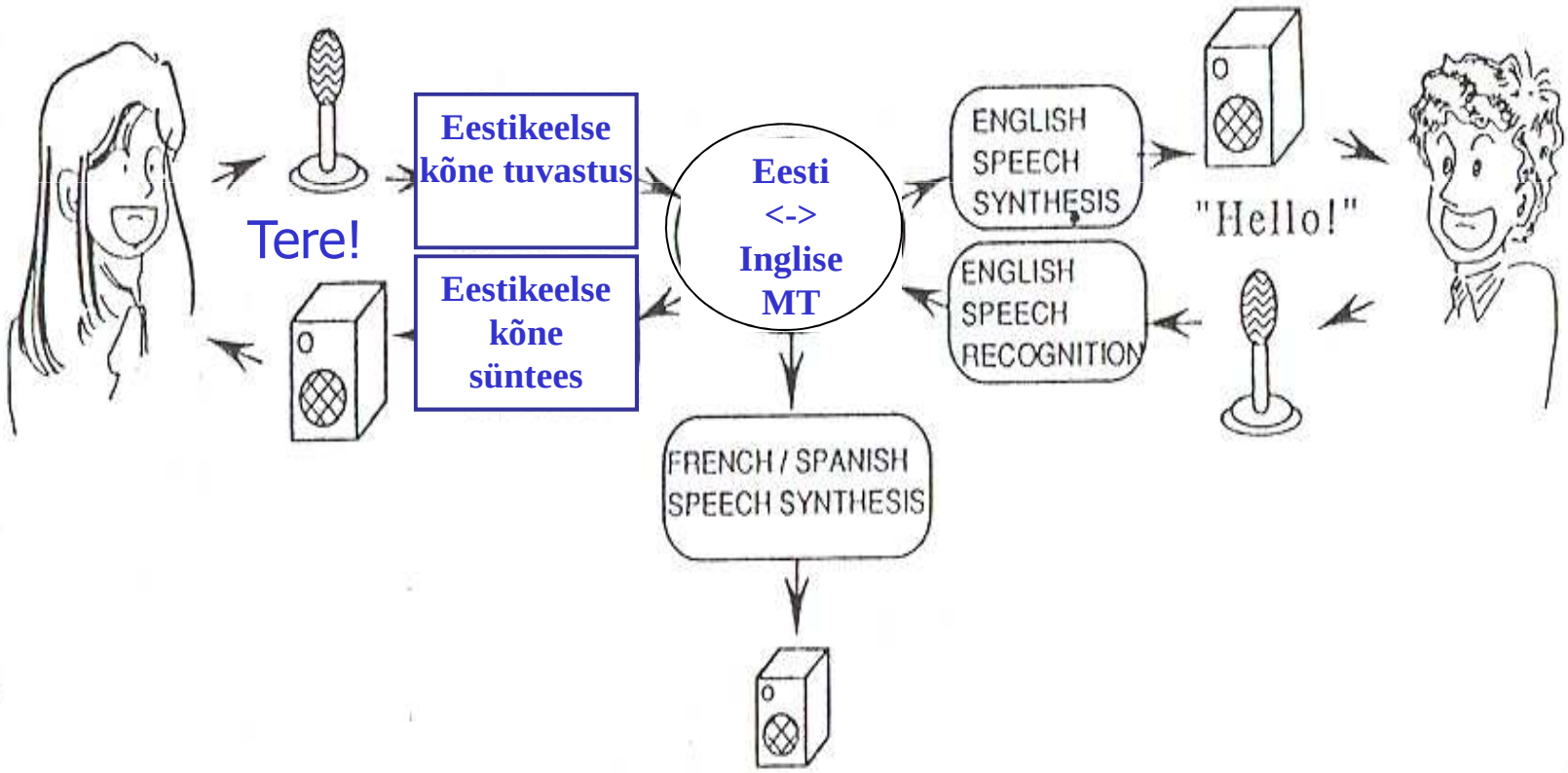
2010: Apple iPad

2010: Google voice search

2011: Apple Siri

2012: Microsoft speech-to-speech translation

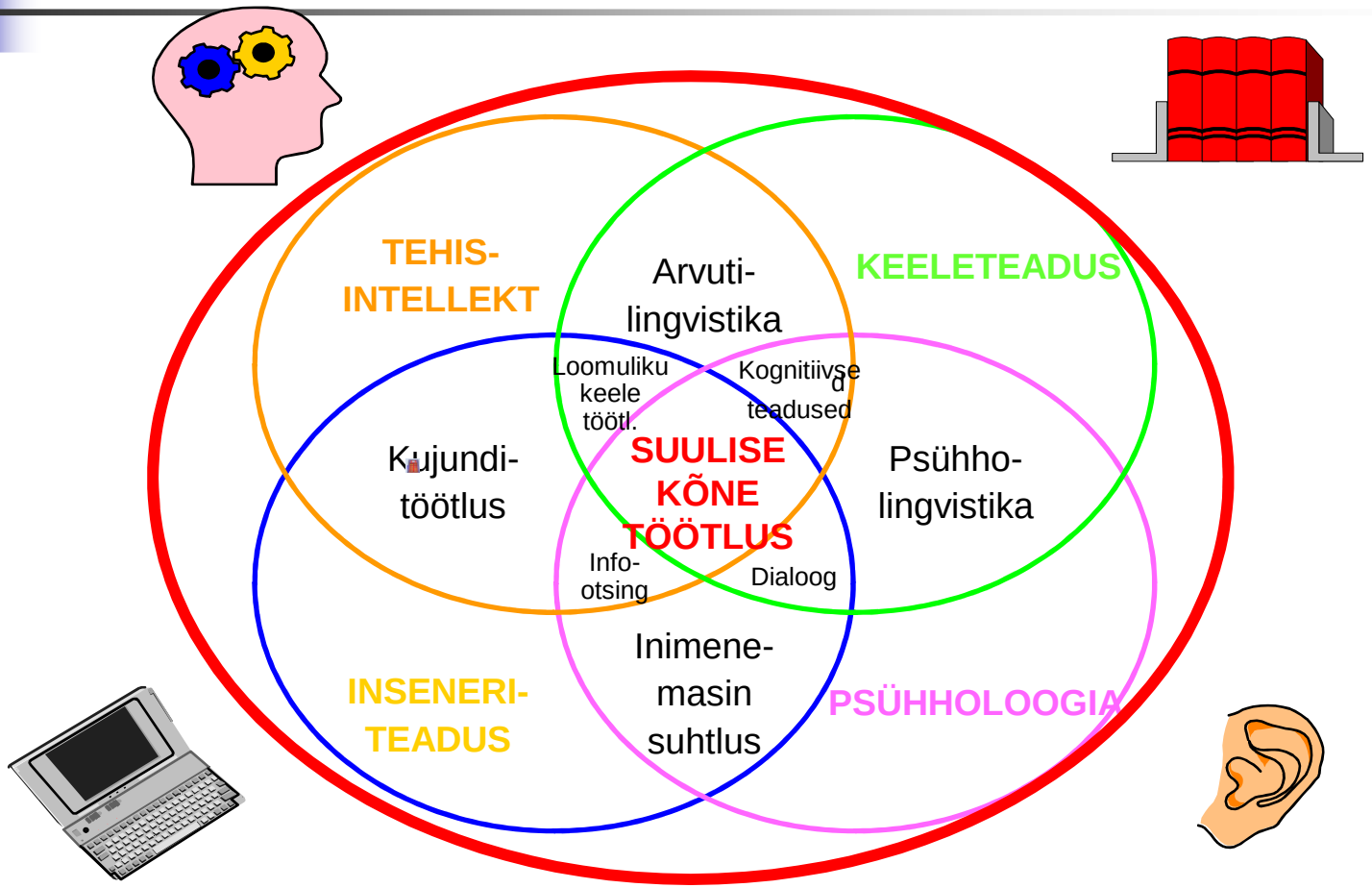
Keeletehnoloogia olemus



Keeletehnoloogia on ...

... keelealaste teadmiste rakendamine selliste arvutisüsteemide loomiseks, mis võimaldavad analüüsida, tuvastada, mõista ja sünteesida inimkeelt

Keeletehnoloogia on interdistsiplinaarne



Keeletehnoloogia komponendid

Tehnoloogilised lahendused:
kõnesüntees ja tuvastus, morfoloogiline, süntaktiline ja semantiline analüüs, masintõlge, keeleõppevahendid, jne

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

Keeled ja tehnoloogia

- 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 masintöötluse tehnoloogia (~ 1-2%)
- Prognoos: 100 aasta pärast on maailmas kasutusel vaid 50-10% täna eksisteerivatest keeltest

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

Mitmekeelsus Euroopa Liidus

- 27 liikmesriiki, 23 ametlikku keelt / 506 keelepaari
- Euroopa Komisjonis töötab 2500 tõlki – aastas tõlgitakse 1,8 miljonit lk
- Täieliku mitmekeelsuse tagamiseks oleks vaja 8500 tõlki

- Kui palju läheb see meile maksma?
- 1,1 miljardit eurot aastas = 2,2 eurot iga EL kodaniku kohta

Mitmekeelsus Euroopa Liidus

- 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

Eesti keel ja tehnoloogia

- Riiklik programm “Eesti keele keeletehnoloogiline tugi (2006-2010)”
- Riiklik programm “Eesti keeletehnoloogia 2011–2017”
- Keeletehnoloogia programmide eesmärk: luua KT arenduseks vajalikud keeleressursid ja keelespetsiifilised tehnoloogilised lahendused

Keeletehnoloogia riiklik programm (2006-2010)

Kokku 24 projekti:

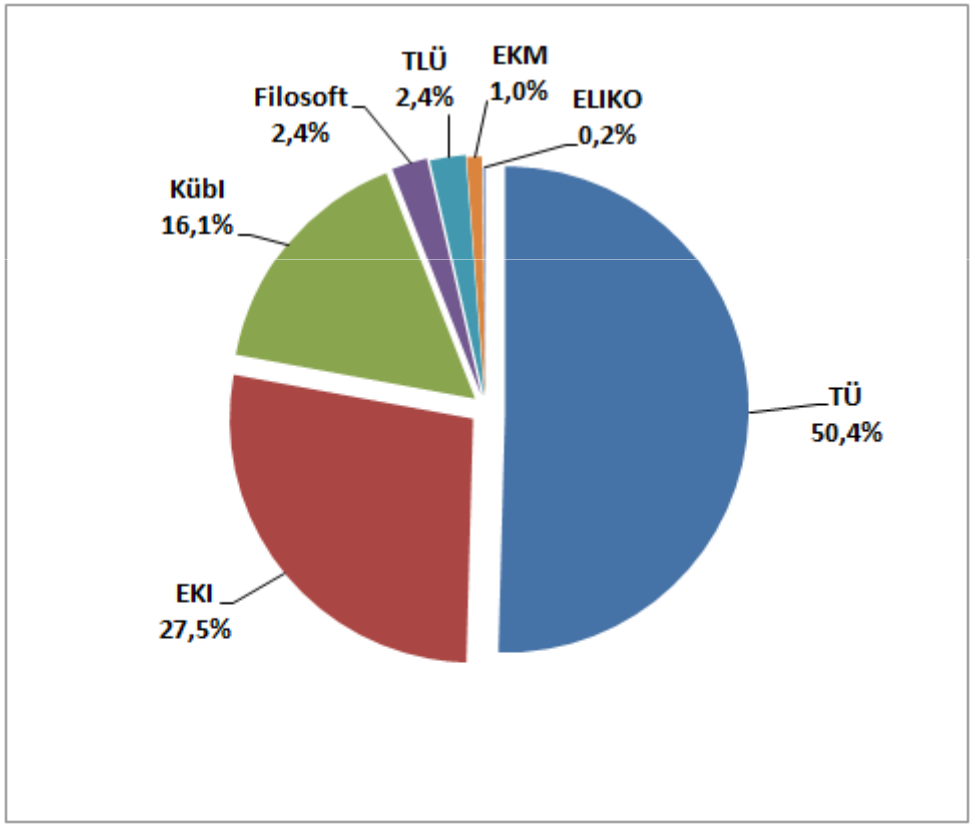
- Kõnekorpused – emotsionaalne kõne, spontaanne kõne, aktsendiga kõne, dialoogid
- Tekstikorpused – kirjakeele korpus, mitmekeelsed paralleelkorpused, korpusepäringud
- Tehnoloogilised lahendused – kõnetuvastus, kõnesüntees, masintõlge, sõnastike töövahendid, infootsing, süntaktiline ja semantiline analüüs, dialoogimudelid
- http://www.keeletehnoloogia.ee/projektid

Finantseerimine 2006–2010

	2006	2007	2008	2009	2010
Projektitaotluste arv	22	22	23	24	24
Finantseeritud projektide arv	18	20	23	23	24
Kogusumma, MEEK (MEUR)	7,3 (0,47)	7,1 (0,46)	13,4 (0,86)	12,9 (0,83)	11,8 (0,75)

Kokku 2006 - 2010: ~ 53 MEEK (~ 3,4 MEUR)

Finantseerimine 2006–2010



Keeletehnoloogia riiklik programm (2011-2017)

Meede 1: **Tarkvaraprototüüpe loovad uurimis- ja arendusprojektid**

10 projekti:

- Kõne ja teksti emotsionaalsuse statistilised mudelid (EKI)
- Kõnesünteesiliidesed (EKI)
- Leksikograafi töökeskkonna modifitseerimine (EKI)
- E-keelenõu (EKI)
- Eestikeelse dialoogi pragmaatika analüsaator (TÜ)
- Vahendid teksti mitmekihiliseks märgendamiseks (TÜ)
- Semantika vahendid eesti keelele (TÜ)
- Mallipõhine faktituletus tekstikorpustest (TÜ)
- Kõnetuvastus (TTÜ KübI)
- Audiovisuaalse kõnesünteesi prototüüp (TTÜ KübI)

Keeletehnoloogia riiklik programm (2011-2017)

Meede 2: **Keeleressursse loovad projektid**

9 projekti:

- Eesti Wordnet'i täiendamine (TÜ)
- Eesti keele spontaanse kõne foneetilise korpuse arendused (TÜ)
- Autentse meditsiinikeele korpuse alusel radioloogia elektroonse piltsõnastiku koostamine (TÜ)
- Suulise eesti keele audiovisuaalse suhtluskorpuse kogumine ja päringusüsteemi arendamine (TÜ)
- Uued ressursid masintõlkes (TÜ)
- Kõne- ja multi-modaalsed korpused (TTÜ KübI)
- Võru ja seto keelekorpus (Võru Instituut)
- Eesti-prantsuse paralleelkorpus (EPLÜ)
- Eesti avatud paralleelkorpus (Tilde Eesti OÜ)

Keeletehnoloogia riiklik programm (2011-2017)

Meede 3: **Eesti Keeleressursside Keskus**

Keskne depositoorium keelekorpuste ja -tarkvara arhiveerimiseks, haldamiseks ja kõigile huvilistele kättesaadavaks tegemiseks

http://www.keeleressursid.ee

EKRK:

- loodud TÜ, TTÜ KübI ja EKI konsortsiumina
- teaduse tuumiktaristu objekt
- üle-euroopalise konsortsiumi CLARIN-ERIC liige

Keeletehnoloogia riiklik programm (2011-2017)

Meede 4:

Integreeritud keeletarkvara ja selle rakendused

2 projekti:

- Subtiitrite helindamise ja tele-eetrisse edastamise tarkvaralahendus (EKI, ERR, EPL)
- Eestikeelsete dialoogsüsteemide loomise raamistik (TÜ, Hambaravikliinik ja TÜ ajaloo muuseum)

Keeletehnoloogia riiklik programm (2011-2017)

Meede 5: **Tellitavad projektid**

Tellija: juhtkomitee või avaliku sektori asutus

2012.a konkurss vabavaralise morfoloogiatarkvara arendamiseks

http://www.keeletehnoloogia.ee/ekt-projektid

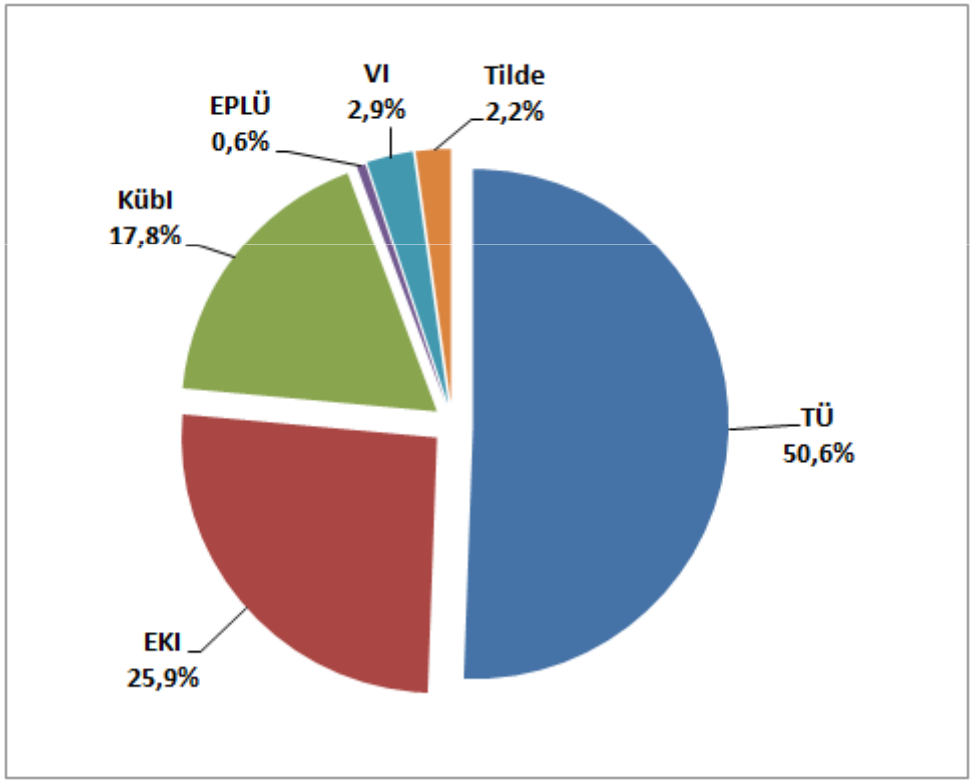
Finantseerimine 2011–2017

	2011	2012	2013	2014	2015	2016	2017
Projektitaotluste arv	21	21	25	27	28	29	30
Finantseeritud projektide arv	18	19	21	22	23	24	25
Kogusumma, MEUR	0,75	0,64	0,72	0,75	0,75	0,75	0,75

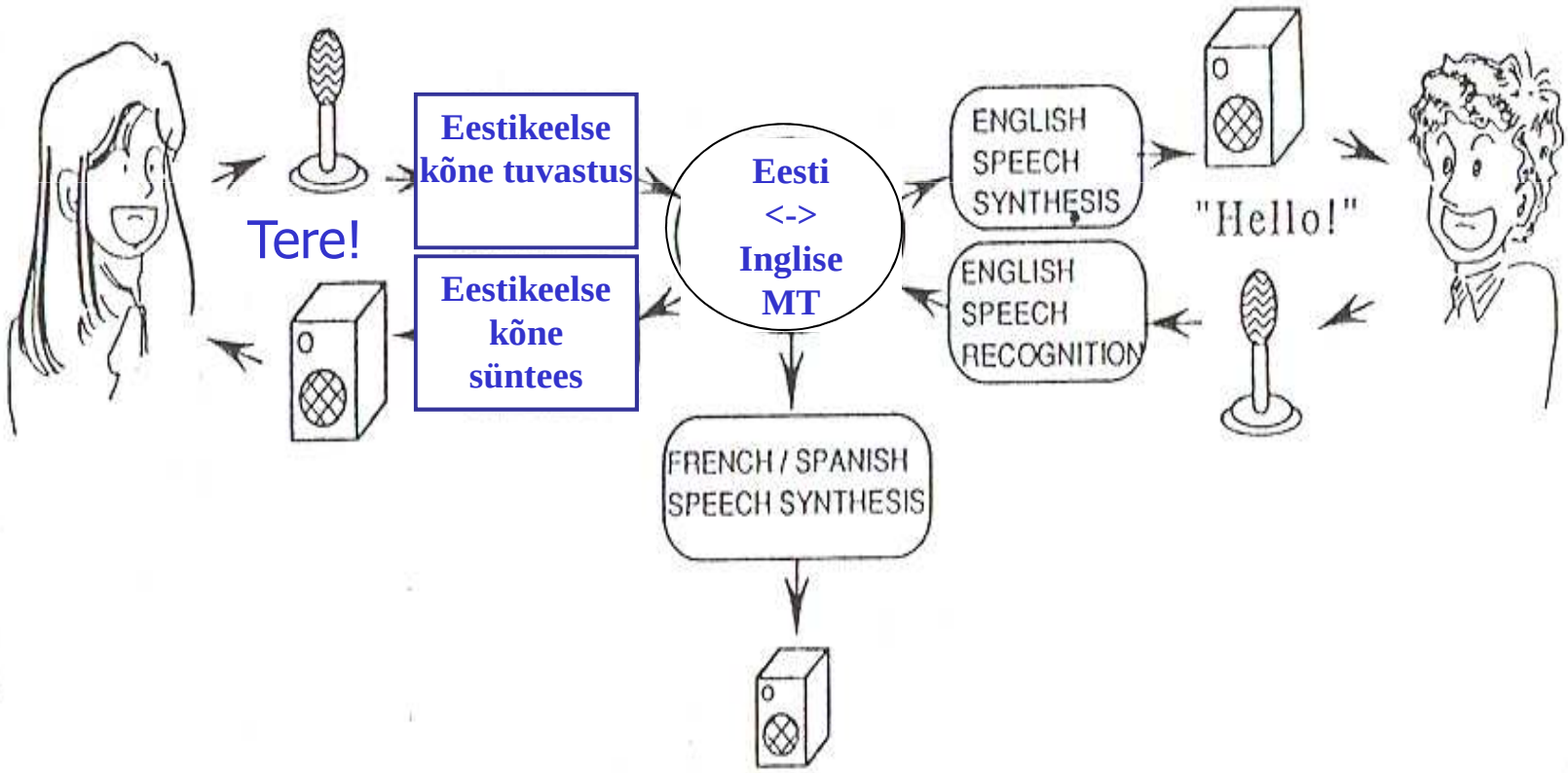
Kokku 2011-2017: ~ 5,1 MEUR

Võrdluseks: Prantsusmaa Quaero-programm (2008-2013): 200 MEUR

Finantseerimine 2011–2013



Reaalsus lähiaastatel (?)



Näiteid prototüüpidest

- Eesti <--> inglise masintõlge

  http://masintolge.ut.ee/
  http://translate.google.com/

Thomas J. Watson: *„Ma arvan, et maailmas on turgu võib-olla ainult viie arvuti jaoks".*

TÜ: *I think the world is maybe the market only five for the computer*

Google: *I think the world is maybe the only market for the five computer*

Näiteid prototüüpidest

- Eesti <--> inglise masintõlge

  http://masintolge.ut.ee/
  http://translate.google.com/

Ken Olson: *„Ma ei näe mingit põhjust, miks inimene peaks endale koju arvutit tahtma".*

TÜ: *I see no reason why the person should get yourself home computer want*

Google: *I do not see any reason why a person should want your home computer*

Näiteid prototüüpidest

- Kõnesüntees – ortograafilise teksti teisendamine inimkõneks
- Sünteesidemo: http://heli.eki.ee/koduleht/ 

Rakendused:

- Veebisõnastike helindamisliides
- Nutitelefoni rakendused
- Heliraamatute genereerija moodulid
- Subtiitrite helindamine

Näiteid prototüüpidest

- Kõnetuvastus – kõne teisendamine kirjalikuks tekstiks
- Rakendused:
  - Pikkade kõnesalvestuste transkribeerimise veebiteenus
  - Reaalajalise kõnetuvastuse veebiteenus
  - Kõnesalvestuste brauser
  - Mobiilirakendused:
    - Kõnele
    - Arvutaja
    - Diktofon
    - Inimesed
  - Kõnetuvastus radioloogidele

Näiteid prototüüpidest

- Pikkade kõnesalvestuste transkribeerimise veebiteenus http://bark.phon.ioc.ee/webtrans/

- Kahekümne neljandal ja kahekümne viiendal aprillil toimub Väike-Maarjas rahvusvaheline konverents millega tähistatakse kahekümne viie aasta möödumist Ferdinand Johann viide mõni keeleauhinna asutamisest toimuvast teeb veebiülekanded Art. ülikooli multimeediakeskus seda saab jälgida kahekümne neljandal aprillil alates kella üheteistkümnest ja kahekümne viiendal aprillil alates kella üheksast.

Näiteid prototüüpidest

- Kõnesalvestuste brauser
  http://bark.phon.ioc.ee/tsab

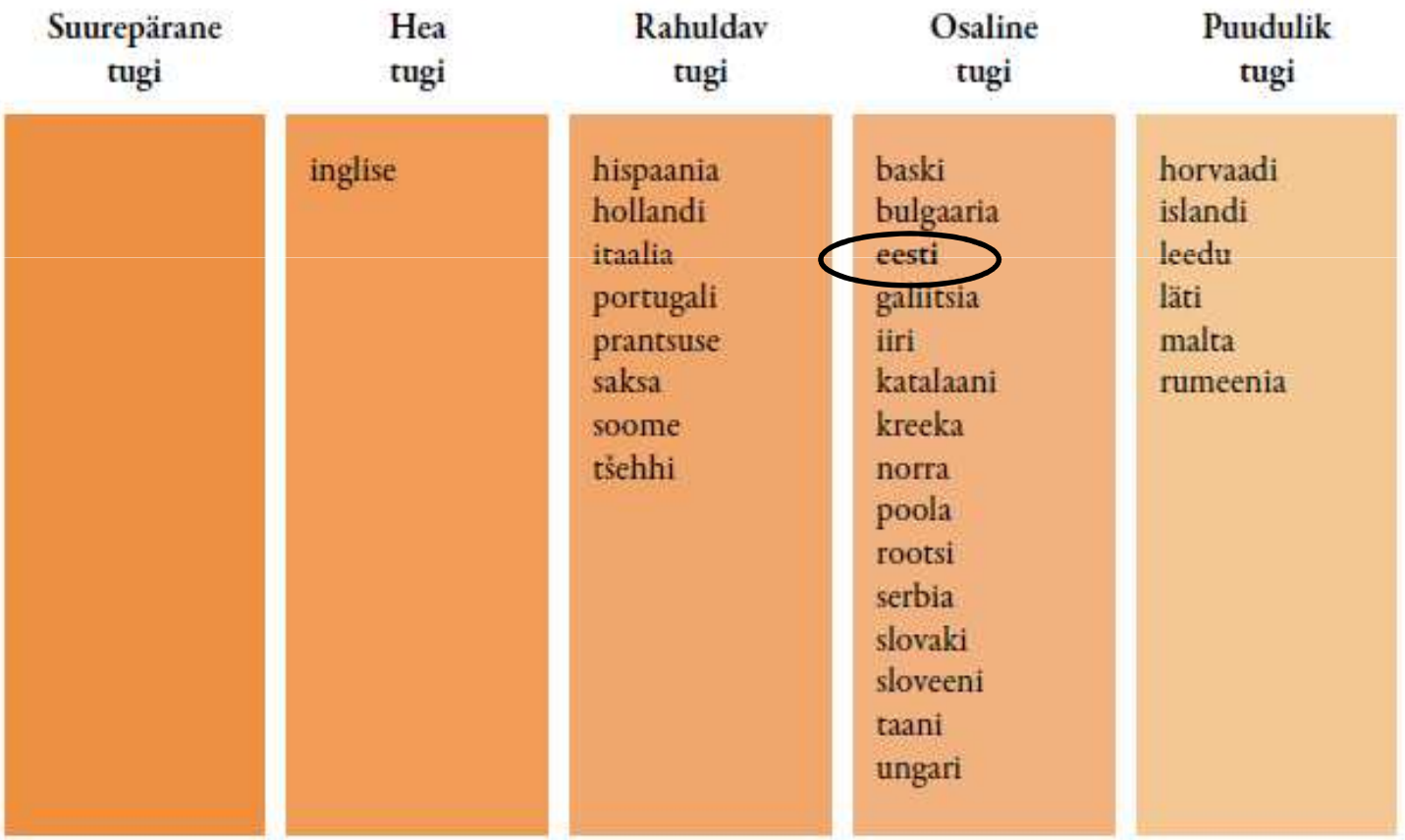
- Üle 3000 automaatselt transkribeeritud raadiosaate aastast 2010
- Tekst ja heli sünkroonis vaadatav/kuulatav
- Otsimisvõimalus saadetest märksõna põhjal

Eesti KT Euroopas

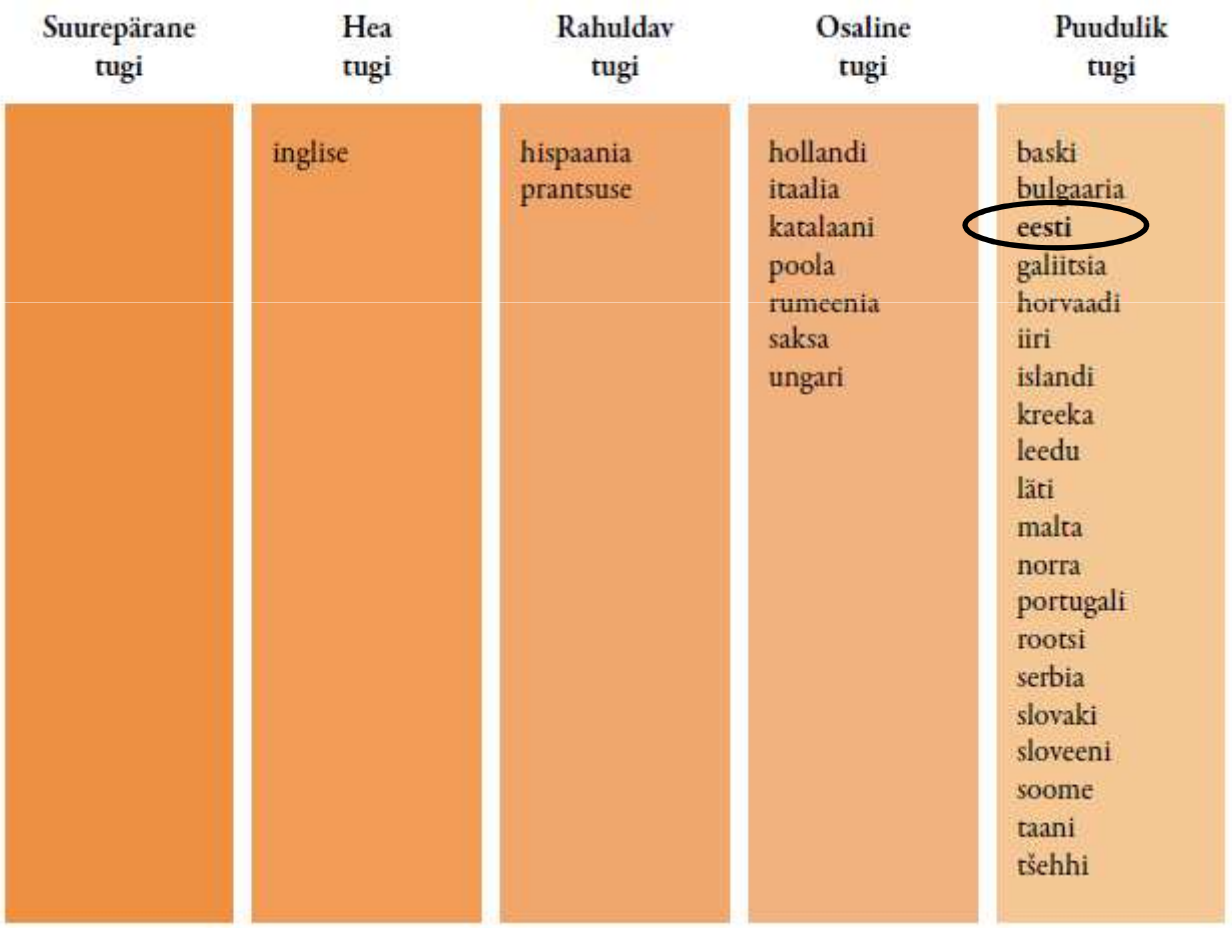
- META-NET: Euroopa põhiliste keelte tehnoloogilise toe arengu võrdlus (2012) http://www.meta-net.eu/whitepapers/overview

- Võrdluses 30 keelt
- Eesti esindaja META-NETis Tartu ülikool

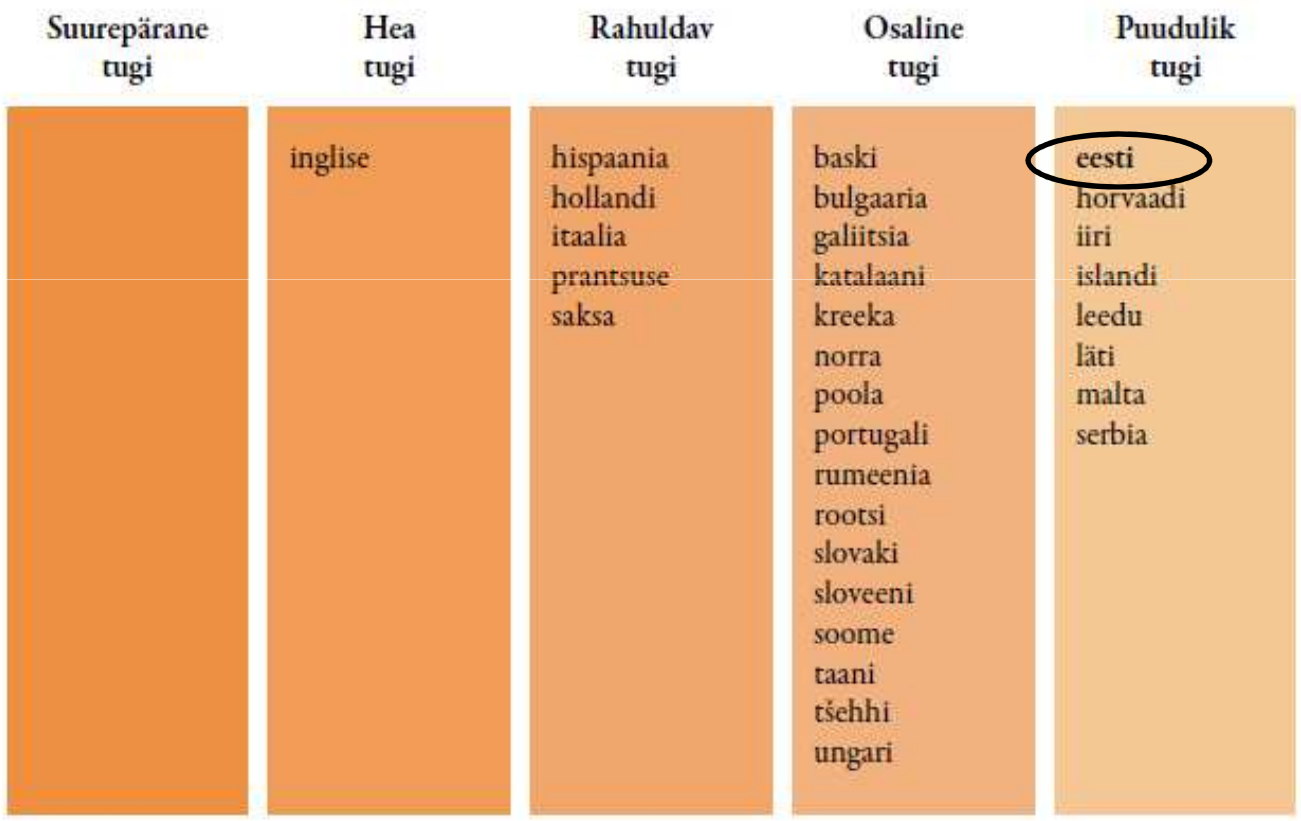
Kõnetöötlus



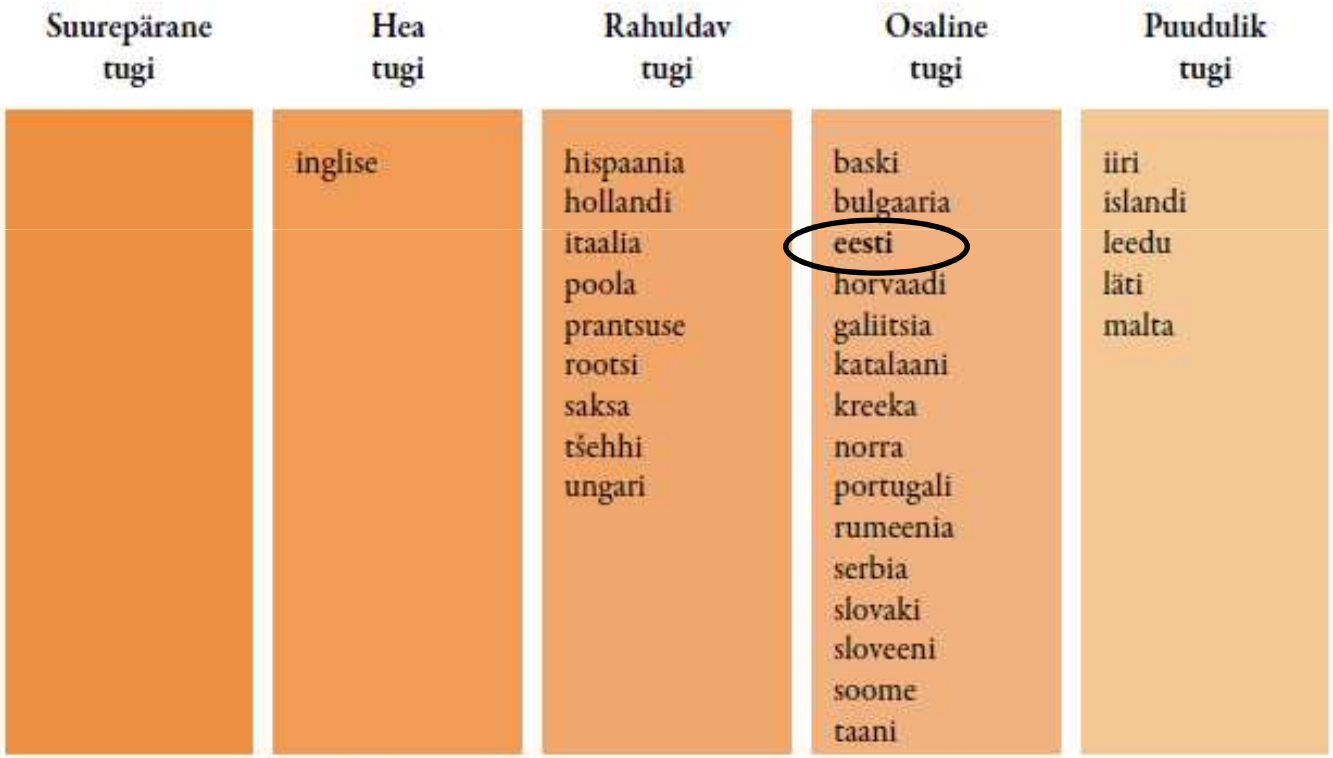
Masintõlge


Suurepärane tugi
Hea tugi
Rahuldav tugi
Osaline tugi
Puudulik tugi
inglise
hispaania
prantsuse
hollandi
itaalia
katalaani
poola
rumeenia
saksa
ungari
baski
bulgaaria
eesti
galiitsia
horvaadi
iiri
islandi
kreeka
leedu
läti
malta
norra
portugali
rootsi
serbia
slovaki
sloveeni
soome
taani
tšehhi

Tekstianalüüs



Kõne- ja tekstikorpused



Tehnoloogilise toe pingerida

Kokkuvõtteks

- META-NETi hinnang eesti keele tehnoloogilisele toele: **„ettevaatlikult optimistlik“**
- Parem kui läti ja leedu keele tehnoloogiline tugi tänu stabiilsele rahastusele
- Keeletehnoloogia RP idee on õige – IT-arendajatel on tekkinud huvi prototüüpide kasutamiseks
- **Rohkem ressurssi on vaja uurijate-arendajate järelkasvu koolitamiseks**
- Keeletehnoloogia ei saa valmis aastal 2017
- Eesti keele digitaalset hääbumist pole lähiaastatel põhjust karta

Lõpetuseks

“Tänu keeletehnoloogia arendamisele on eesti keel elujõuline ka saja aasta pärast!”