META NET

White Paper Series Valge raamatu sari

THE ESTONIAN LANGUAGE IN THE DIGITAL AGE

EESTI KEEL DIGIAJASTUL

Krista Liin
Kadri Muischnek
Kaili Müürisep
Kadri Vider

Springer

White Paper Series Valge raamatu sari

THE ESTONIAN LANGUAGE IN THE DIGITAL AGE

EESTI KEEL DIGIAJASTUL

Krista Liin Tartu Ülikool
Kadri Muischnek Tartu Ülikool
Kaili Müürisep Tartu Ülikool
Kadri Vider Tartu Ülikool

Georg Rehm, Hans Uszkoreit
(toimetajad, editors)

EESSÕNA PREFACE

Eesti keele raport kuulub META-NETi väljaannete sarja, mille eesmärgiks on tutvustada keeletehnoloogia-alaseid teadmisi ja selle ala potentsiaali. Väljaande sihtgrupiks on õpetajad, ajakirjanikud, poliitikud, kogu keelekogukond ja teised teemast huvitatud.

Keeletehnoloogia kättesaadavus ja kasutamine on Euroopa keeliti väga erinev. Nii on ka meetmed, mida on vaja rakendada keeletehnoloogia arendamise ja uurimise edasiseks toetamiseks, erinevatele keeltele väga erinevad, sõltudes näiteks keele keerukusest ja selle kõnelejate arvust.

Euroopa Komisjoni rahastatud tippteadmiste võrgustik META-NET viis läbi keeleressursside ja -tehnoloogiate alase uurimuse, mis keskendus 23 ametlikule Euroopa keelele ja ka teistele olulistele regionaalsetele keeltele Euroopas (vt lk 75). Analüüsi tulemus näitas, et kõigi keelte tehnoloogiates leidub märkimisväärseid puudujääke. Täpne ekspertanalüüs ja olukorra hindamine aitavad panustada edasise uurimistöö mõju suurendamisse ja vähendada riske.

META-NET koosneb 33 riigi 54 uurimiskeskusest (vt lk 71), mis teevad koostööd tööstuse, valitsusasutuste, ülikoolide ja uurimisasutuste esindajatega. Koostöö tulemusena valmib ühine tehnoloogiline visioon, mis osana strateegilisest uurimiskavast näitab, kuidas keeletehnoloogilised rakendused saavad katta praegused uurimistöö puudujäägid aastaks 2020.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 75). The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

As of January 2012, META-NET consists of 54 research centres from 33 European countries (p. 71). META-NET is working with stakeholders from economy (software companies, technology providers, users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Selle dokumendi autorid tänavad saksa keele valge raamatu autoreid loa eest kasutada nende väljaandes sisaldunud keelest sõltumatuid materjale [1].

Selle keeleraporti koostamist rahastas 7. raamprogramm ja Euroopa Komisjoni IKT poliitika toetusprogramm lepingute T4ME (toetusleping 249 119), CESAR (toetusleping 271 022), METANET4U (toetusleping 270 893) ja META-NORD (toetusleping 270 899) kaudu.

The authors of this document are grateful to the authors of the White Paper on German for permission to re-use selected language-independent materials from their document [1].

The development of this White Paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

SISUKORD CONTENTS

EESTI KEEL DIGIAJASTUL

1 Kokkuvõte 1

2 Oht meie keeltele ja väljakutse keeletehnoloogiale 4
- 2.1 Keelepiirid tõkestavad Euroopa infoühiskonda . . . 5
- 2.2 Meie keeled on ohus . . . 5
- 2.3 Keeletehnoloogia on võtmetehnoloogia . . . 5
- 2.4 Keeletehnoloogia võimalused . . . 6
- 2.5 Keeletehnoloogia väljakutsed . . . 7
- 2.6 Kuidas inimesed ja masinad keelt omandavad . . . 7

3 Eesti keel Euroopa infoühiskonnas 9
- 3.1 Üldinfo . . . 9
- 3.2 Eesti keele eripärad . . . 9
- 3.3 Viimase aja arengud . . . 10
- 3.4 Keelehoole Eestis . . . 11
- 3.5 Keel ja haridus . . . 12
- 3.6 Rahvusvahelised aspektid . . . 12
- 3.7 Eesti keel internetis . . . 12

4 Eesti keele keeletehnoloogiline tugi 14
- 4.1 Rakenduste arhitektuur . . . 14
- 4.2 Kesksed rakendused . . . 15
- 4.3 Muud rakendusalad . . . 22
- 4.4 Haridusprogrammid . . . 24
- 4.5 Riiklikud programmid ja algatused . . . 24
- 4.6 Vahendite ja ressursside kättesaadavus . . . 25
- 4.7 Keeltevaheline võrdlus . . . 26
- 4.8 Järeldused . . . 28

5 META-NETist 31

THE ESTONIAN LANGUAGE IN THE DIGITAL AGE

1 Executive Summary 33

2 Languages at Risk: a Challenge for Language Technology 36
- 2.1 Language Borders Hold back the European Information Society . . . 37
- 2.2 Our Languages at Risk . . . 37
- 2.3 Language Technology is a Key Enabling Technology . . . 38
- 2.4 Opportunities for Language Technology . . . 38
- 2.5 Challenges Facing Language Technology . . . 39
- 2.6 Language Acquisition in Humans and Machines . . . 39

3 The Estonian Language in the European Information Society 41
- 3.1 General Facts . . . 41
- 3.2 Particularities of the Estonian Language . . . 41
- 3.3 Recent Developments . . . 43
- 3.4 Language Cultivation in Estonia . . . 43
- 3.5 Language in Education . . . 44
- 3.6 International Aspects . . . 45
- 3.7 Estonian on the Internet . . . 45

4 Language Technology Support for Estonian 47
- 4.1 Application Architectures . . . 47
- 4.2 Core Application Areas . . . 48
- 4.3 Other Application Areas . . . 56
- 4.4 Educational Programmes . . . 57
- 4.5 National Programmes and Initiatives . . . 58
- 4.6 Availability of Tools and Resources . . . 59
- 4.7 Cross-language comparison . . . 60
- 4.8 Conclusions . . . 61

5 About META-NET 65

A Kirjandus – References 67

B META-NETi liikmed – META-NET Members 71

C META-NETi Valge raamatu sari – The META-NET White Paper Series 75

1

KOKKUVÕTE

Viimase 60 aasta jooksul on Euroopas välja kujunenud küll ühtne poliitiline ja majanduslik struktuur, kuid kultuuri ja keelte osas on mitmekesisus säilinud. Keelelised takistused pärsivad nii Euroopa kodanike omavahelist kui ka äri- ja poliitikaringkondade suhtlust erinevates keeltes - portugali keelest poola keeleni ja kreeka keelest keldi keeleni. Euroopa Liidu asutused kulutavad aastas miljoneid eurosid mitmekeelsuspoliitika tagamiseks, s.t tõlgitakse tekste ja suulisi vestlusi. Aga kas meil oleks võimalik neid kulutusi vältida? Tänapäeva keeletehnoloogia ja keeleteadus annavad suure panuse keelebarjääri lõhkumiseks. Tulevikus aitab keeletehnoloogia koos nutikate seadmete ja programmidega eurooplastel üksteisega suhelda ja äri ajada isegi siis, kui nad ei räägi sama keelt.

Keeletehnoloogia ehitab sillad Euroopa tulevikku.

Üks võimalus (kuid seejuures mõeldamatu võimalus) Euroopa mitmekeelsuse probleemi lahendamiseks oleks kasutusele võtta üks domineeriv keel ja sellega teised keeled asendada.
Klassikaline moodus keelebarjääri ületamiseks on võõrkeelte õppimine. Ent tehnilise toeta on majanduse, poliitväitluste ja teadusarenduse tarbeks kõigi Euroopa Liidu 23 ametliku liikmesriigi keele ja 60 muu Euroopa keele omandamine kodanikele ületamatu takistus.
Lahenduseks on võtmetehnoloogiate välja arendamine. Digitaalne keeletehnoloogia hõlmab kõiki kirjaliku ja suulise keele suhtluse vorme. Seega soodustab ta koostööd, äritegevust, teadmiste jagamist ning ühiskondlikus ja poliitilises diskussioonis osalemist, sõltumata seejuures kasutaja võimalikust keelebarjäärist ja arvutikasutamise oskuse tasemest. Sageli on keeletehnoloogia juba keerulistesse süsteemidesse lõimitud. Tulevikus võiks keeletehnoloogilistest lahendustest moodustuda ainulaadne Euroopa keelte vaheline sild.
Eesmärgi saavutamiseks ja samas Euroopa kultuurilise ja keelelise mitmekesisuse säilitamiseks tuleb esmalt süstemaatiliselt analüüsida iga Euroopa keele lingvistilist eripära ja seda toetava keeletehnoloogia hetkeseisu.
Eesti keelt kõneleb emakeelena umbes miljon inimest ja see on Eesti Vabariigi ainuke ametlik keel. Eesti keele igapäevast kasutust reguleerib keeleseadus. Samas on Eesti tuntud e-valitsuse ja e-riigi poliitika poolest. Eesti keel teaduse ja kõrghariduse keelena tugineb pikaajalisele eestikeelse kõrghariduse ja teadustöö traditsioonile. Erinevalt enamusest Euroopa keeltest ei kuulu eesti keel indoeuroopa keelkonda. Eesti keele eripäradeks võib lugeda täishäälikute rohkust, täis- ja kaashäälikute kolme pikkust, artiklite ja grammatilise soo puudumist. Samuti on eesti keelele iseloomulik rikkalik muutemorfoloogia. Eesti keele liitsõnamoodustus on vaba ja produktiivne. Sõnajärg lauses on küllaltki vaba.

Keeletehnoloogia kui võti tulevikku.

Praegu turul kättesaadavad automaattõlke- ja kõnetöötlusvahendid selle eesmärgini veel ei küündi. Põhilised turul tegutsejad on kasumi saamisele suunatud Põhja-Ameerika eraettevõtted. 1970ndatel hakati Euroopa Liidus tähtsustama keeletehnoloogiat kui Eu-

roopat ühendavat jõudu ja samal ajal alustati ka riiklike projektidega, mis andsid küll väärtuslikke tulemusi, kuid ei aidanud kaasa Euroopa ühistegevusele. Tänu mitmete varasemate ja jätkuvate teadus- ja arendustöö programmide toetusele on keeletehnoloogiline uurimismaastik Eestis olemas.

Inimkeele keerukus raskendab loomuliku keele modelleerimist tarkvaras ning rakenduse tegelikus elukeskkonnas testimine on pikk ja kulukas protsess. Kahjuks ei ole näiteks inglise keelele arendatud keelemudelid eesti keelele ülekantavad, sest eesti keelel on vabam sõnajärg, peaaegu piiranguteta liitsõnade moodustamine ning suurem käände- ja pöördelõppude hulk. Ometi on aastatepikkuse töö tulemusena loodud töökindel eesti keele õigekirjakontroll (speller), mis on lõimitud ka levinumatesse kontoritarkvara pakettidesse.

Eestikeelne infootsing Google otsimootoriga on veebikasutajate seas niivõrd levinud, et 2009. aastast alates on sõna guugeldama lisatud ka Eesti Õigekeelsussõnaraamatusse. Keelest sõltumatud otsinguvahendid suudavad leida ainult sõnavorme, millel on päringusõnaga täpselt sama kuju või mis sisaldavad päringusõna alamsõnena. Kuid kuna eesti keele morfoloogia on rikas ja lisaks lõppudele võib ka sõna tüvi muutuda, siis on edukaks otsinguks ja indekseerimiseks vaja keelespetsiifilisi vahendeid. Keelespetsiifilised indekseerijad leiavad enne sõnade indeksisse lisamist nende algvormid ehk lemmatiseerivad otsisõnad. Eesti Infosüsteemide Amet on avalikult soovitanud kasutada Eesti avaliku sektori infosüsteemide infootsingul ja indekseerimisel lemmatiseerimismoodulit [2].

Kaks peamist keeletehnoloogiasüsteemides kasutatavat meetodit "omandavad" keelelised oskused inimestega sarnasel viisil. Statistilised ehk andmejuhitud meetodid omandavad keelelise teadmuse suurtest näidistekstide kogudest. Teine meetod on reeglipõhiste süsteemide loomine, mille suureks eeliseks on asjaolu, et ekspertidel on keele töötluse üle täpsem kontroll. Toetudes senistele tähelepanekutele, näib, et tänapäeva "hübriidne" keeletehnoloogia, mis ühendab keele süvatöötluse statistiliste meetoditega, suudab ületada kõigi Euroopa ja muudegi keelte vahelise lõhe.

Keeletehnoloogia valdkonnas on Euroopa teadustöö olnud edukas. Näiteks kasutatakse Euroopa Liidu tõlketeenustes avatud lähtekoodiga masintõlke tarkvara Moses, mida arendati peamiselt Euroopa teadusprojektide raames. Eesti keele masintõlge on tõsine väljakutse. Sõnastikupõhise analüüsi muudab keeruliseks vaba liitsõnamoodustus, uusi sõnu saab liitmise teel alati juurde tekitada. Analüüsiprobleeme põhjustavad ka vaba sõnajärg ja mitmeosalised tegusõnad (ühend- ning väljendverbid). Lisaks kõigele muule on piiratud ka paralleelsete tekstide hulk. Vaatamata sellele kuulub Eesti keel nende ligi 50 maailma keele hulka, mida saab arvuti abil tõlkida.

Tulevikus on oodata märkimisväärseid muutusi kõnetehnoloogia arengus. Juba praegu pakutakse Eestis nutitelefonide kasutajatele tsentraliseeritud teenustena kõne dikteerimist. Sarnased TTÜ Küberneetika Instituudis välja töötatud eestikeelsed kõnetuvastusrakendused nutitelefonidele võitsid 2011. aasta parima keeleteo auhinna.

Käesolev keeleraportite sari näitab, et Euroopa Liidu liikmesriikides on keeletehnoloogilised lahendused ja teadustöö erineval tasemel. Tõeliselt efektiivsete tehnoloogiliste lahendusteni jõudmiseks vajavad põhjalikumat uurimistööd veel isegi Euroopa suurimad keeled, rääkimata eesti keele keeletehnoloogia arendamisest.

Eesti keele keeletehnoloogilise olukorra hinnang annab põhjust ettevaatlikuks optimismiks. Eesti keele jaoks on olemas nii kõnetuvastuse kui ka -sünteesi vahendid. Nende edasine arendustöö on hetkel aktiivselt käimas. Vaatamata eesti keele keerulisele morfoloogiale, on eesti keele morfoloogiaanalüsaatori efektiivsus võrreldav teiste Euroopa keelte vastavate vahenditega, kuid süntaksianalüsaatoritel on veel palju arenguruumi.

Keele genereerimise vahenditest on olemas ainult morfoloogilise sünteesi programmid. Laiem üldsus kasutab masintõlkeks Google'i tõlketeenust, Tartu Ülikoolis on arendamisel ka eesti-inglise masintõlkesüsteem. Ilmselt oleks suur nõudlus ka eesti-vene-eesti masintõlkele. Enamik neist vahenditest on loodud uurimisasutustes ja neid võib pidada pigem prototüüpideks, mitte valmis toodeteks. Kahjuks esindavad Eesti keeletehnoloogiatööstust ainult mõned üksikud väikeettevõtted nagu Filosoft. Viimastel kümnenditel on loodud märkimisväärne hulk Eesti keele ressursse (korpused, leksikonid, WordNet), seega olukord keelelise andmestiku osas on küllaltki hea.

Keeletehnoloogia aitab Euroopat ühendada.

Mis puutub keerukamatesse valdkondadesse nagu tekstisemantika, keele genereerimine ja märgendatud multimodaalsed ressursid, siis eesti keele jaoks põhivahendid ja -ressursid puuduvad. Eesti keele keeletehnoloogilist uurimistööd ja arendustegevust on toetanud mitmed riiklikud keeletehnoloogia-alased uurimisprogrammid, seetõttu on nii loodud ressursid kui vahendid vabaks kasutamiseks.

Käesolev keeleraportite sari täiendab teisi META-NETi strateegilisi tegevusi (ülevaade on saadaval raporti lisas). META-NETi kodulehelt http://www.meta-net.eu leiab uuemat informatsiooni, näiteks META-NETi visiooni [3] või strateegilise uurimiskava (SRA) uusima versiooni. META-NETi pika-ajalisem eesmärk on võimaldada kõigile keeltele kõrgekvaliteedilist keeletehnoloogiat ja kultuurilise mitmekesisuse kaudu saavutada poliitiline ja majanduslik ühtsus.

2

OHT MEIE KEELTELE JA VÄLJAKUTSE KEELETEHNOLOOGIALE

Oleme tunnistajateks digirevolutsioonile, mis avaldab tohutut mõju meie suhtlusele ja ühiskonnale. Viimast arengut digitaalses info- ja kommunikatsioonitehnoloogias võrreldakse Gutenbergi trükipressi leiutamise mõjuga. Mida ütleb see analoogia meile Euroopa infoühiskonna, täpsemalt meie keelte tuleviku kohta?

> Me oleme tunnistajaks digitaalsele revolutsioonile, mis on võrreldav Gutenbergi trükipressi leiutamisega.

Pärast Gutenbergi leiutist toimus tõeline läbimurre kommunikatsioonis ja teadmiste jagamises, näiteks tõlkis Luther Piibli rahvakeelde. Sellele järgnenud sajanditel on arendatud kultuuritehnoloogiaid keeletöötluse ja teadmistevahetuse edendamiseks:

- suuremate keelte õigekirja ja grammatika standardiseerimine tegi võimalikuks teaduse ja ideede kiire leviku;
- ametlike keelte areng võimaldas kodanikel teatud (sageli poliitiliste) piiride raames suhelda;
- keelte õpetamine ja tõlkimine tegi võimalikuks keelteülese suhtluse;
- kirjutiste toimetamise ja bibliograafiaalaste juhtnööride loomine kindlustas trükimaterjalide kvaliteedi ja kättesaadavuse;
- uut liiki meedia – ajalehtede, raadio, televisiooni, raamatute ja muude formaatide – teke rahuldas erinevaid kommunikatsioonivajadusi;

Viimase kahekümne aasta jooksul on infotehnoloogia aidanud kaasa mitme protsessi automatiseerimisele ja lihtsustamisele, nt:

- kirjastustarkvara on asendanud masinakirja ja trükiladumise;
- Microsoft PowerPoint on asendanud lüümikud ja grafoprojektorid;
- meilidega saadetakse ja saadakse dokumente kiiremini kui faksi teel;
- Skype annab võimaluse odavateks internetikõnedeks ja virtuaalsete koosolekute pidamiseks;
- audio- ja videokodeeringud lihtsustavad multimeedia jagamist;
- otsingumootorid lubavad veebilehtedeni jõuda märksõnade kaudu;
- veebiteenused, nagu näiteks Google Translate, annavad kiireid ligikaudseid tõlkeid;
- sotsiaalmeedia platvormid, näiteks Facebook, Twitter ja Google+, lihtsustavad suhtlust, koostööd ja infovahetust.

Kuigi neist tööriistadest ja rakendustest on abi, ei suuda need veel toetada jätkusuutlikku mitmekeelset Euroopa ühiskonda, kus info ja kaup liiguksid vabalt.

2.1 KEELEPIIRID TÕKESTAVAD EUROOPA INFOÜHISKONDA

Me ei oska täpselt ennustada, milline näeb välja tuleviku infoühiskond. Kuid on väga tõenäoline, et kommunikatsioonitehnoloogia revolutsioon ühendab uuel moel eri keeli kõnelevaid inimesi. See paneb inimesed uusi keeli õppima ja arendajad looma uusi rakendusi, mis aitaksid kaasa üksteisemõistmisele ja võimaldaksid juurdepääsu jagatud teadmisele. Uued meedialiigid seovad üha rohkem keeli, kõnelejaid ja teavet, mis liigub ülemaailmses majandus- ja infosfääris. Sotsiaalmeedia (Wikipedia, Facebook, Twitter, YouTube, viimasel ajal ka Google+) praegune populaarsus on vaid jäämäe tipp.

Tänapäeval saame saata gigabaitides teksti ümber maailma kõigest paari sekundiga, enne kui taipame, et see oli kirjutatud keeles, mida me ei mõista. Euroopa Komisjoni hiljutise uuringu kohaselt ostab 57% internetikasutajatest Euroopas tooteid ja teenuseid keeltes, mis ei ole nende emakeel. Kõige levinum võõrkeel on inglise keel, sellele järgnevad prantsuse, saksa ja hispaania keel. 55% kasutajatest loeb võõrkeelseid materjale, samas kui vaid 35% kasutab teist keelt ise meilide kirjutamisel või veebikommentaaride postitamisel [4]. Mõned aastad tagasi oli inglise keel interneti *lingua franca* – valdav enamus veebist oli inglisekeelne – ent praeguseks on olukord drastiliselt muutunud. Teistes Euroopa keeltes (aga ka Aasia ja Lähis-Ida keeltes) oleva materjali maht on internetis plahvatuslikult kasvanud.

Üllataval kombel pole see keelepiiridest tulenev üldlevinud digitaalne lõhe pälvinud kuigi suurt avalikkuse tähelepanu. Samas tõstatab see pakilise küsimuse: milliseid Euroopa keeli saadab võrgupõhises info- ja teadmusühiskonnas edu ja millised on määratud kaduma?

> Maailmamajandus ja inforuum seavad meid vastamisi erinevate keelte, kõnelejate ja sisuga.

2.2 MEIE KEELED ON OHUS

Kuigi trükipress aitas kaasa Euroopasisese infovahetuse kiirenemisele, viis see ka paljud Euroopa keeled väljasuremiseni. Piirkondlikke ja vähemuskeeli trükiti harva, nii säilisid näiteks korni ja dalmaatsia keel vaid suulisel kujul, see omakorda piiras oluliselt nende kasutusvaldkonda. Kas interneti mõju meie keeltele on samasugune?

Euroopa ligi 80 keelt on üks tema väärtuslikumaid ja tähtsamaid kultuuriväärtusi ning eluline osa tema ainulaadsest ühiskonnamudelist [5]. Samal ajal kui inglise või hispaania keelel pole tõenäoliselt probleeme tekkival digitaalsel turul ellujäämisega, võivad mitmed Euroopa keeled võrguühiskonnas vähetähtsaks jääda. See omakorda aga nõrgestaks kogu Euroopa positsiooni maailmas ja oleks vastuolus meie strateegilise eesmärgiga kindlustada võrdsed võimalused kõigile Euroopa kodanikele, olenemata nende emakeelest.

> Euroopa keeleline mitmekesisus on meie üks rikkamaid ja olulisimaid kultuurivarasid.

UNESCO mitmekeelsuse raporti järgi on keeled hädavajalik vahend oma põhiõiguste, näiteks poliitilise väljendusvabaduse, hariduse ja ühiskonnas osalemise tagamiseks [6].

2.3 KEELETEHNOLOOGIA ON VÕTMETEHNOLOOGIA

Varem tähendas keele säilitamine keeleõppele ja tõlkele keskendumist. Arvatakse, et 2008. aastal oli tõlkimise, tarkvara lokaliseerimise ja veebilehtede globaliseerimise turuosa Euroopas 8,4 miljardit eurot, ning ennustatakse, et see kasvab 10% aastas [7]. Samas katab see summa vaid väikese osa praegusest ja tulevasest keeltevahelisest kommunikatsioonivajadusest. Ahvatlev la-

hendus tagamaks tuleviku Euroopas keelekasutuse laia katvust ja head kvaliteeti oleks keeletehnoloogia kasutamine, samamoodi nagu me kasutame tehnoloogiat transpordi- ja energiavajaduste rahuldamiseks.

Digitaalne keeletehnoloogia hõlmab kõiki kirjaliku ja suulise keele suhtluse vorme. Seega soodustab ta koostööd, äritegevust, teadmiste jagamist ning ühiskondlikus ja poliitilises diskussioonis osalemist, sõltumata seejuures kasutaja võimalikust keelebarjäärist ja arvutikasutamise oskuse tasemest. Sageli on keeletehnoloogia juba keerulistesse süsteemidesse lõimitud ja see aitab meil:

- otsimootori abil veebist informatsiooni leida;
- tekstiredaktoriga õigekirja ja grammatikat kontrollida;
- veebipoes tootesoovitusi näha;
- auto navisüsteemi hääljuhiseid kuulda;
- internetiteenuste abil veebilehti tõlkida.

Keeletehnoloogia koosneb mitmetest kesksetest rakendustest, mis suuremas rakenduste raamistikus on vajalikud teiste programmide tööks. META-NETi keeleraportite eesmärgiks välja selgitada iga Euroopa keele tuumikrakenduste tase.

Euroopa vajab veakindlat ja kättesaadavat keeletehnoloogiat kõigi Euroopa keelte jaoks.

Jätkuvalt ülemaailmselt innovatiivseks eeskujuks olemiseks vajab Euroopa kõigile oma keeltele kohandatud keeletehnoloogiat, mis oleks nii robustne (veakindel) kui taskukohane ja samas olulisematesse IT-süsteemidesse tihedalt lõimitud. Lähitulevikus ei jõuta ilma keeletehnoloogiata mitmekeelse ning tõeliselt efektiivse ja interaktiivse multimeediapõhise kasutajakogemuseni.

2.4 KEELETEHNOLOOGIA VÕIMALUSED

Trükitehnika läbimurdeks oli võimalus teksti (lehekülge) trükipressi abil kiiresti kopeerida. Teadmiste otsimise, lugemise, tõlkimise ja kokkuvõtmise raske töö jäi inimestele. Kõne salvestamiseks tuli oodata Edisoni – ja ka tema tehnoloogia suutis luua kõigest analoogkoopiaid. Kaasaegne keeletehnoloogia võimaldab automatiseerida kõigis Euroopa keeltes tõlkimise, sisutootmise ja teadmushalduse. Tänu sellele on võimalik luua koduelektroonikale, masinatele, sõidukitele, arvutitele ja robotitele intuitiivseid keelel ja kõnel põhinevaid kasutajaliideseid. Reaalselt kasutatavad äri- ja tööstusrakendused on praegu alles arendamise algusjärgus. Kuid saavutused teadusvallas on tekitanud rakenduste loomiseks uusi võimalusi. Nii näiteks töötab masintõlge kindla valdkonna raames juba mõistliku täpsusega ning on olemas eksperimentaalseid rakendusi, mis pakuvad mitmekeelset infot, teadmushaldust ning sisutootmist paljudes Euroopa keeltes.

Nagu teistegi tehnoloogiatega, loodi ka esimesed keeletehnoloogia rakendused (kõnepõhised kasutajaliidesed ja dialoogisüsteemid) kindlatele valdkondadele ning seetõttu oli nende efektiivsus sageli piiratud. Tohutu turupotentsiaaliga on haridus- ja meelelahutustööstus. Keeletehnoloogiat lõimitakse mängudesse, harivasse meelelahutusse, raamatukogudesse, simulatsioonidesse ja treeningprogrammidesse. Keeletehnoloogia mängib olulist rolli mobiilsetes infoteenustes, arvutipõhises keeleõppetarkvaras, e-õppe keskkonnas, enesehindamisprogrammides, plagiaatide tuvastamise tarkvaras ning paljudes teistes rakendusvaldkondades. Twitteri- ja Facebookilaadsete sotsiaalmeediarakenduste populaarsusega kaasneb suurenenud vajadus keeletehnoloogia järele, mis peaks jälgima postitusi, võtma kokku arutelusid, hindama arvamustrende, leidma emotsionaalseid vastuseid, tuvastama ja jälitama autoriõiguse rikkumisi ja väärkasutust.

Keeletehnoloogia loob Euroopa Liidule tohutuid võimalusi. See aitab lahendada keerulisi mitmekeelsuse probleeme, mis tekivad Euroopa ettevõtetes, asutustes ja koolides erinevate keelte koos kasutamise tõttu. Keeletehnoloogia võimaldab kodanike suhtlust Euroopa ühisturul, kõrvaldades takistavad keelebarjäärid, ent samas toetades üksikute keelte vaba kasutust.

Keeletehnoloogia aitab saada üle keelelise mitmekesisuse "puudest".

Tulevikus on Euroopa innovaatiline mitmekeelne keeletehnoloogia eeskujuks meie ülemaailmsetele partneritele, kui nad alustavad oma mitmekeelsete kogukondade toetamisega. Keeletehnoloogiat võib pidada tugitehnoloogiaks, mis aitab jagu saada keelelise mitmekesisuse "puudest" ja muudab keelekogukonnad üksteisele lihtsamini ligipääsetavateks.

Lõpuks veel ühest aktuaalsest uurimisvaldkonnast – keeletehnoloogia kasutamisest katastroofipiirkondade päästeoperatsioonidel. Kriisiolukorras tegutsemine võib olla elu ja surma küsimus, seega keelest sõltumatute oskustega intelligentsed robotid suudaksid päästa elusid.

2.5 KEELETEHNOLOOGIA VÄLJAKUTSED

Kuigi viimastel aastatel on keeletehnoloogia märkimisväärselt arenenud, on praegune tehnoloogiline edasiminek ja tooteinnovatsioon siiski liiga aeglased. Laialdaselt kasutatavad tehnoloogiad, nagu tekstiredaktorite spellerid ja grammatikakorrektorid, on tüüpiliselt ükskeelsed ja saadaval vaid loetud keeltele.

Praegune tehnoloogilise arengu tempo on liiga aeglane.

Veebipõhised masintõlketeenused on küll kasulikud dokumendi sisust kiire ülevaate saamiseks, ent nad jäävad hätta täpse ja täieliku tõlkega. Inimkeele keerukus raskendab loomuliku keele modelleerimist tarkvaras ning rakenduse tegelikus elukeskkonnas testimine on pikk ja kulukas protsess, mis vajab järjepidevat rahalist toetust. Selleks, et Euroopa oleks endiselt mitmekeelse kogukonna tehnoloogia teerajaja rollis, tuleb leiutada uusi meetodeid arengu kiirendamiseks. Need hõlmavad nii tarkvaralisi uuendusi kui *crowdsourcingu* stiilis tehnikaid.

2.6 KUIDAS INIMESED JA MASINAD KEELT OMANDAVAD

Et näitlikustada, kuidas arvutid keelt käsitlevad ja miks on nii raske arvuteid loomuliku keele kasutamiseks programmeerida, anname lühikese ülevaate sellest, kuidas inimesed keelt omandavad ning kuidas keeletehnoloogiasüsteemid töötavad.

Inimesed omandavad keeleoskuse kahel viisil: õppides näidetest ja õppides keelereegleid.

Inimesed omandavad keeli kahel erineval viisil. Väikelapsed omandavad emakeele vanemate, õdede-vendade ja teiste pereliikmete vahelist suhtlust kuulates. Umbes teisel eluaastal lausuvad lapsed oma esimesi sõnu ja lühikesi fraase. Keeleõpe on võimalik tänu inimeste geneetilisele soodumusele kuuldut imiteerida ja mõtestada.

Vanemas eas nõuab teise keele omandamine suuremat pingutust, peamiselt seetõttu, et õppija ei kuulu emakeelena kõnelejate kogukonda. Koolis õpitakse võõrkeeletundides tavaliselt selgeks keele grammatiline struktuur, sõnavara ja õigekiri. Õppimiseks kasutatakse harjutusi, mis kirjeldavad keelelist teadmust abstraktsete reeglite,

tabelite ja näidete abil. Vanemaks saades muutub võõrkeele omandamine raskemaks.

Kaks peamist keeletehnoloogiasüsteemides kasutatavat meetodit "omandavad" keelelised oskused sarnasel viisil. Statistilised ehk andmejuhitud meetodid omandavad keelelise teadmuse suurtest näidistekstide kogudest. Kui näiteks spelleri treenimiseks piisab ükskeelsetest tekstidest, siis masintõlkesüsteemi treenimiseks läheb vaja paralleeltekste kahes või enamas keeles. Treeningtekstidest "õpib" masintõlkealgoritm sõnade, fraaside ja lausete tõlkimiseks mustreid.

Selline statistiline lähenemine vajab toimimiseks miljoneid lauseid. Mida rohkem näitetekste analüüsitakse, seda parem tõlketulemus saadakse. Tekstiredaktorites olev speller ning näiteks Google'i otsingumootor ja tõlge kasutavad statistilist lähenemist. Andmejuhitud meetodi eeliseks on see, et masin õpib järjestikustes treeningtsüklites kiiresti, kuigi tulemuse kvaliteet võib oluliselt varieeruda.

Teine meetod, mida keeletehnoloogias ja kitsamalt ka masintõlkes kasutatakse, on reeglipõhiste süsteemide loomine. Keeleteaduse, arvutuslingvistika ja arvutiteaduse valdkonna eksperdid kodeerivad esmalt grammatilised analüüsid (tõlkereeglid) ja koostavad sõnade nimestikud (leksikonid). See on vägagi aeganõudev ja töömahukas tegevus. Mõnda juhtivat tõlkesüsteemi on pidevalt arendatud juba üle kahekümne aasta. Reeglipõhiste süsteemide suureks eeliseks on asjaolu, et eksperdil on keele töötluse üle täpsem kontroll. See teeb võimalikuks tarkvaras leiduvate vigade süstemaatilise parandamise ja kasutajale täpsema tagasiside andmise, seda eriti siis, kui reeglipõhised süsteemid on kasutuses keeleõppe abina. Kõrge kulu tõttu on seni reeglipõhiseid süsteeme arendatud üksnes suuremate keelte jaoks.

Keeletehnoloogiasüsteemide kaks peamist tüüpi omandavad keelt samal viisil.

Kuna statistiliste ja reeglipõhiste süsteemide plussid ja miinused kalduvad teineteist täiendama, siis uuemad uurimused keskenduvad neid lähenemisi kombineerivatele hübriidsüsteemidele. Kahjuks pole need süsteemid seni tööstusrakendustes sama edukad olnud kui teadituslaborites.

Käesolevast peatükist selgus, et paljud tänapäeva infoühiskonnas laialt levinud rakendused on tihedalt seotud keeletehnoloogiaga. Võttes arvesse meie mitmekeelset kogukonda, kehtib see väide iseäranis selgelt Euroopa majandus- ja infosfääri puhul. Kuigi keeletehnoloogia on viimastel aastatel märkimisväärselt arenenud, on veel kõvasti arenguruumi süsteemide kvaliteedi parandamise osas.

Järgnevalt toome välja eesti keele rolli Euroopa infoühiskonnas ja hindame eesti keele keeletehnoloogilise toe praegust seisu.

3

EESTI KEEL EUROOPA INFOÜHISKONNAS

3.1 ÜLDINFO

Eesti keelt kõneleb emakeelena umbes miljon inimest. Peamiselt räägitakse seda Eestis (922 000 kõnelejat), aga ligi 160 000 eesti keele kõnelejat kasutab seda ka Venemaal, Ameerika Ühendriikides, Rootsis, Kanadas, Soomes ja mitmetes teistes maades [8]. 2000. aasta rahvaloenduse andmetel on Eestis 1 370 052 elanikku, kellest 167 804 kõnelevad eesti keelt võõrkeelena [9]. Eesti keel on Eesti Vabariigi ainuke ametlik keel.

> Eesti keelt kõneleb emakeelena umbes miljon inimest.

Eesti keele variantide hulka kuuluvad eesti keele piirkondlikud variandid (murded ja nende kirjakeeled, erinevates välisriikides kõneldavad keelevariandid), erinevate ühiskonnagruppide keelevariandid - sotsiolektid ning keelealaste erivajadustega inimeste keelevariandid (sh. viipekeel).

Eesti keele piirkondlike variantide alla kuuluvad eesti murded ja nende kirjakeeled. Kõige suuremad erinevused on Põhja-Eesti ja Lõuna-Eesti murrete vahel. Need keeleerinevused on pärit juba meie ajaarvamise eelsest ajast, mil Uurali keelte läänemeresoome harust hakkasid eristuma iseseisvad keeled. Asjaolu, et siinsed elanikud elasid kuni 19. sajandi lõpuni väga paikset elu, aitas kaasa piirkondlike murrete tekkele; eristatakse kuni sadat kohalikku murrakut. Tänapäeva eesti keel arenes välja Põhja-Eesti murrete põhjal, toetudes osaliselt ka Lõuna-Eesti murrakutele [10].

Tänapäeval kõneldakse murdekeelt peamiselt Lõuna-Eestis ja läänepoolsetel saartel. Võru ja setu murded väärivad eraldi mainimist kui standardsest kirjakeelest kõige erinevamad. Riik toetab eesti keele piirkondlike variantide kasutamist ja nende säilitamist kultuuriväärtusena, kirjakeele allikana ning kohalike eestlaste identiteedi kandjatena. Paljudes koolides Võru- ja Viljandimaal õpetatakse kohalikke keeli (vastavalt võru, setu ja mulgi keelt) valikainena.

Väliseesti keel on eesti keele variant, õigemini küll variandid, mida räägivad püsivalt väljaspool Eestit elavad keelekõnelejad esimese või teise keelena. Mõnel juhul on Eestist väljarännanute emakeel säilinud ja iseseisvalt arenenud rohkem kui sajandi vältel. Loomulikult mõjutavad neid variante tugevalt asukohamaal kõneldavad keeled. Ligi 2000 Eestis elava kurdi emakeeleks või peamiseks suhtlusvahendiks on eesti viipekeel (õigemini eesti viipekeel ja viibeldud eesti keel), mida kasutavad ka kuulmispuudega eestlased ning kurtide ja kuulmispuudega inimeste hooldajad [11].

3.2 EESTI KEELE ERIPÄRAD

Eesti keel kuulub Uurali keelkonna läänemeresoome harusse koos soome, karjala ja muude lähisugulaskeeltega. Eesti keel on kaugemalt sugulane ka ungari keelega. Oluline aspekt on see, et erinevalt enamusest Euroopa keeltest ei kuulu Uurali keeled indoeuroopa keelkonda.

Tüpoloogiliselt esindab eesti keel üleminekuvormi aglutineerivalt keelelt fusiivsele keelele. Läbi ajaloo on talle avaldanud suurt mõju saksa keel, seda nii sõnavara kui süntaksi osas.

Eesti keele eripäradeks võib lugeda rõhu esinemist esimesel silbil, täishäälikute rohkust, kolme eristatavat pikkust täis- ja kaashäälikutel (välted), artiklite ja grammatilise soo puudumist (ka asesõnades) ning indoeuroopa keeltest erinevat baassõnavara. Samuti on eesti keelele iseloomulik rikkalik muutemorfoloogia: käändsõnad muutuvad 14 käändes ja kahes arvus, pöördsõnad ajas, isikus, kõneviisis, tegumoes ja kõneliigis.

Kuigi eesti keeles on 14 käänet, ei kuulu sinna hulka akusatiivi – sihitis võib kontekstist olenevalt esineda nii osastavas, omastavas kui nimetavas käändes. Eesti keele liitsõnamoodustus on vaba ja produktiivne, nn juhuliitsõnu moodustatakse vastavalt vajadusele ja järelikult ei ole kõiki tekstides esinevaid liitsõnu võimalik sõnaraamatus üles lugeda. Teine produktiivne sõnamoodustusviis on tuletamine.

Erinevalt enamusest Euroopa keeltest ei kuulu eesti keel indoeuroopa keelkonda.

Eesti keeles ei ole grammatilist aega tuleviku jaoks ja tulevikus toimuvat väljendatakse sageli tegusõnaga olevikus, tegevuse toimumisaeg selgub kontekstist.

Ta saabub homme.

Euroopa keeltega võrreldes on küllaltki erilised ka eesti keele tingiv ja kaudne kõneviis. Tingiva kõneviisi tunnuseks on liide *-ks(i)-*, sellega väljendatakse hüpoteetilist olukorda või ebamäärast/ebakindlat olukorda.

Kui ta treeniks rohkem, jookseks ta kiiremini.

Kaudse kõneviisi tunnuseks on tegusõna lõpus olev *-vat*. Selle kõneviisiga väljendatakse sündmusi, millest teatakse kuulu järgi.

Ta jooksvat kiiresti.

Kuigi eesti keelt on kategoriseeritud SVO keeleks, on sõnajärg küllaltki vaba, kusjuures tüüpiliselt asub verb lauses teisel kohal. Sõnajärge mõjutab lause infostruktuur – tuntud ja uue informatsiooni eristamine.

- *Ta jooksis kiiresti koju.*
- *Kiiresti jooksis ta koju.*
- *Koju jooksis ta kiiresti.*
- *Jooksis ta kiiresti koju?*
- *Kui ta kiiresti koju jooksis, siis ...*

Kuigi eesti keel on lähedane soome keelele, on pikaajaline saksa keele mõju seda oluliselt muutnud ja lähendanud nn keskmisele Euroopa keelele (Standard Average European, SAE) [12]. Soome keelest erinevate SAE-päraste joontena võiks nimetada sõnajärge teatud kõrvallausetüüpides või ühendverbide rohket kasutust üldse ja eriti aspekti (tegevuse lõpetatuse) väljendamiseks, vrd eesti *Ta tegi selle ära* ja soome *Hän teki sen*. Samuti on eesti keeles tunduvalt rohkem võõrsõnu ja hiliseid laensõnu kui soome keeles.

Eesti keele ortograafia aluseks on foneetiline ehk häälduslähedusе põhimõte, mille järgi taotletakse õigekirja võimalikult head vastavust häälдusele. Eesti keele kirjapanekuks kasutatakse ladina tähestikku, mille baasvariandile on lisatud tähed õ, ä, ö ja ü, võõrsõnades kasutakse ka tähti š ja ž.

Eestikeelne lugeja leiab ülevaate eesti keele struktuurist ning õigekeelsusnormidest Mati Erelti, Tiiu Erelti ja Kristiina Rossi "Eesti keele käsiraamatust" [13]. Inglise-keelsele lugejale võiks soovitada Mati Erelti toimetatud teost "Estonian Language" [14].

3.3 VIIMASE AJA ARENGUD

Eesti keelt on mõjutanud saksa (alguses keskalamsaksa, hiljem saksa kirjakeel), vene ja inglise keel, kuigi ükski neist pole eesti keelega suguluses.

Pärast Teist Maailmasõda viidi Eestis läbi venestamine. Alates iseseisvuse saavutamisest aastal 1918 riigikeeleks olnud eesti keele tähtust vähendati. Pärast Nõukogude Liidu kokkuvarisemist aastal 1991 sai eesti keel jälle ainsaks riigikeeleks.

Paljudele teistele keeltele tuntud probleemid on saanud ohuks ka eesti keelele: väheneb emakeelsete kõnelejate arv, hägustuvad keelenormid, võõrkeelte tugev mõju, eriti ingliskeelsete suhtlusvõrgustike ja ingliskeelse laiatarbekultuuri pealetung.

Eesti keel, sarnaselt näiteks islandi keelele, on üks väiksemaid keeli maailmas, mis toimib ametliku keelena selle kõigis kasutusaspektides: administratiivkeelena, meedias, kirjanduses, teatris, ettevõtluses, koolides, ülikoolides, teaduses ja mujal.

Viimastel aastakümnetel, pärast Eesti iseseisvumist, on ühest küljest eesti keele positsioon paranenud: eesti keelel on riigikeele staatus ja tema püsimine on tagatud seadustega. Teisalt on aga üleilmastumise ja infoühiskonna arengu tulemusena eesti keele osatähtsus vähenenud. Paljudele teistele keeltele tuntud probleemid on saanud ohuks ka eesti keelele: väheneb emakeelsete kõnelejate arv, hägustuvad keelenormid, võõrkeelte tugev mõju, eriti ingliskeelsete suhtlusvõrgustike ja ingliskeelse laiatarbekultuuri pealetung. Keeletehnoloogia alal on raske suuremate keeltega sammu pidada.

Eesti keele kaitseks on loodud mitu riiklikku organisatsiooni. Keeleinspektsioon hoiab silma peal keeleseaduse täitmisel. Haridus- ja teadusministeeriumi keeleosakond planeerib Eesti keelepoliitikat ja hoolitseb meie keele maailmale tutvustamise eest. Ministeeriumi haldusalas olev Eesti Keelenõukogu on koostanud "Eesti keele arengukava".

3.4 KEELEHOOLE EESTIS

Põhiseaduse kohaselt on Eesti Vabariigi riigikeeleks eesti keel ja riigi kohus on tagada eesti rahvuse, keele ja kultuuri säilimine läbi aegade. Eesti keele säilitamiseks ja arenguks vajalikud meetmed on sätestatud "Eesti keele arendamise strateegias (2004–2010)" [10] ja valmivas "Eesti keele arengukavas (2011–2017)" [15]. Eesti keele igapäevast kasutust reguleerib keeleseadus ja sellel põhinev seadusandlus.

Eesti keele igapäevast kasutust reguleerib keeleseadus ja sellel põhinev seadusandlus.

Eesti keele (ja teiste keelte) arengu ja kasutusega seotud tegevusi koordineerib Haridus- ja teadusministeerium. Eesti keelenõukogu jälgib ja analüüsib Eesti keelelolukorda ning koostab keelestrateegia seiret ja jätkustrateegiaid. Haridus- ja teadusministeeriumi osakondadest tegelevad keeleküsimustega lisaks keeleosakonnale ka Riiklik Eksami- ja Kvalifikatsioonikeskus ja Keeleinspektsioon. Ministeeriumi hallatavatest üksustest tegeleb nende küsimustega Eesti Keele Instituut. Keelekorraldusega tegelevad veel Emakeele Seltsi keeletoimkond, Tartu keelehooldekeskus ning Tartu ja Tallinna ülikoolide õppejõud.

Eesti keel on üks Euroopa Liidu ametlikke keeli, eesti EL terminoloogia areng toimub koostöös Eesti Keele Instituudi terminoloogiaosakonnaga ning Eesti Terminoloogia Ühinguga.

2003. aastal koostas Eesti Keelenõukogu eesti keele arendamise strateegia aastateks 2004–2010, mis sisaldas eesti keele olukorra, seatud eesmärkide ja nende saavutamiseks vajalikke sammude ja asutuste teaduspõhist kirjeldust [10]. Esimene eesti keele arendamise strateegia oli planeeritud katma kõiki peamisi keelekasutuse valdkondi, sealhulgas ka keeletehnoloogiat.

Järgmine eesti keele arendamise strateegia koostati Eesti Keelenõukogu poolt aastal 2010 [15]. "Eesti keele aren-

gukava 2011–2017" on dokument, mis paneb paika peamised strateegilised suunad eesti keele arenguks, õpetamiseks, uurimiseks ja kaitseks. Koos oma rakenduskava, vastavate seadusandlike dokumentide ja muude toetavate tegevustega (nt. rahastamine) kindlustab eesti keele arengukava eesti keele staatuse riigikeelena ja selle jätkuva positsiooni Eesti Vabariigi peamise suhtluskeelena.

3.5 KEEL JA HARIDUS

Haridus on üks tähtsamaid vahendeid keele arengu ja stabiilse positsiooni tagamiseks. Üks hariduse ülesandeid on tagada üldine ja erialane kirjaoskus ning luua mitte-eestlastes positiivne hoiak eesti keele suhtes. Üldharidus, iseäranis kohustuslik üldharidus, on äärmiselt tähtis, sest just see mõjutab keelekasutust kõige rohkem. Seaduse järgi võib põhiharidust omandada ükskõik millises keeles. Praegu kasutatakse gümnaasiumides kahte õppekeelt: kolmveerand koolidest õpib eesti, veerand vene keeles. Eesmärgiga parandada eesti keele oskust mitte-eestlastest gümnaasiumilõpetajate seas alustati 2007. aastal muukeelsetes keskkoolides üleminekuprotsessiga, kus osasid aineid õpetatakse eesti keeles.

Eesti keel on kõigis põhikoolides ja gümnaasiumides (sh vastava taseme haridust andvates kutsekoolides) kohustuslik õppeaine. 2009/2010 õppeaastal oli eestikeelsetes põhikoolides 90 837 õpilast (neist u. 84 000 rahvuselt eestlased), keskharidust andvates õppeasutustes oli see arv 23 769 (neist 22 741 eesti rahvusest) [15].

> Eesti keel teaduse ja kõrghariduse keelena tugineb pikaajalisele eestikeelse kõrghariduse ja teadustöö traditsioonile.

Eesti keel teaduse ja kõrghariduse keelena tugineb pikaajalisele eestikeelse kõrghariduse ja teadustöö traditsioonile. Samas on ülikoolide rahvusvahelistumine toonud kaasa nii võõrkeelse õppe osakaalu suurenemise kui ka välismaalt pärit tudengite ja õppejõudude arvu kasvu. Eesti ülikoolides on pea kõiki erialasid võimalik õppida eesti keeles. Bakalaureuseõppes saab tudeng peaaegu alati omandada oma eriala eesti keeles, kuigi mõnda erialaspetsiifilist ainet võidakse õpetada ka mõnes muus keeles. Siiski on teaduse rahvusvahelistumise tõttu olemas erialakeelte taandumise ja populaarteaduse tasemele jäämise oht - paljudel teadusaladel kirjutatakse ka Eestis kõik doktoritööd ja muud arvestatavad teaduspublikatsioonid inglise keeles.

Mitte-eestlastest täiskasvanute jaoks korraldatakse eesti keele kursusi peamiselt suurema suhtlusvajadusega ametite (meditsiiniõed, politseinikud) esindajatele ja neile, kes taotlevad Eesti kodakondsust (edukatele õppijatele kompenseeritakse õpingukulud). Samuti korraldatakse eesti keele kursusi telesaadetena.

3.6 RAHVUSVAHELISED ASPEKTID

Eesti keel on kuulunud Euroopa Liidu ametlike keelte hulka 2004. aastast alates. See tähendab, et eesti keelt saab kasutada rahvusvahelise suhtluse keelena.

Eesti muutub turistide seas järjest populaarsemaks. Samuti on viimastel aastatel suurenenud eesti keele ja kultuuri vastu huvi tundvate inimeste arv.

Eesti riik toetab eesti keele õpetamist välismaal – hetkel on üle 30 ülikooli, mis pakuvad eesti keele õpet erineval tasemel [16].

3.7 EESTI KEEL INTERNETIS

Statistikaameti andmetel oli Eestis 2010. a ligi 381 300 perekonnal kodune internetiühendus ja 758 100 inimest (55% elanikkonnast) kasutab internetti regulaarselt [17].

Eesti on tuntud e-valitsuse ja e-riigi poliitika poolest. E-riigi poliitika koosneb kahest osast: ühelt poolt interneti

kaudu toimuvad valitsustegevused (valimised, riigi valitsemises osalemine) ja teiselt poolt ligipääs avalikele teenustele. Eesti kodanikud saavad interneti teel näiteks valimistel hääli anda, makse deklareerida, arstiaegu kinni panna ja isegi jälgida oma lapse edasijõudmist koolis.

Eesti on tuntud e-valitsuse ja e-riigi poliitika poolest.

Enamuse siinsete ettevõtete kodulehed on eestikeelsed, ajalehtedel ja -kirjadel on oma uudiste edastamiseks veebiportaalid (http://postimees.ee, http://ohtuleht.ee, http://paevaleht.ee jpm) [18]. On palju teemapõhiseid internetifoorumeid, kus kasutajad suhtlevad eesti keeles. Suhtlusportaalid nagu Orkut ja Facebook on eesti keelde lokaliseeritud. Lisaks leidub palju jututubasid, milles sageli suheldakse kirjakeele normidele mittevastavas keeles – kirjalikus slängis. Vikipeediasse on vabatahtlikud lisanud üle 88 900 eestikeelse artikli.

Keeletehnoloogia vaatepunktist on interneti suurenev osatähtsus oluline kahest aspektist. Ühest küljest kujutab see suur hulk digitaalselt kättesaadavaid keeleandmeid endast rikkalikku materjali loomuliku keele töötluseks, eriti statistilise info kogumiseks. Teisest küljest pakub internet laialdaselt erinevaid võimalusi keeletehnoloogia rakenduseks.

Enim kasutatav veebirakendus on kahtlemata otsingumootor, mis sisaldab keele automaattöötlust erinevatel tasemetel, nagu käesoleva raporti teises pooles täpsemalt võib lugeda. Otsingumootor hõlmab arenenud keeletehnoloogiat, sealjuures iga keele jaoks erinevalt.

Nii Eestis kui mujal Euroopas on välja öeldud, et üheks meie poliitiliseks eesmärgiks on kõigile võrdsete võimaluste tagamine. Avalikel asutustel on kohustus kindlustada puuetega inimestele piiranguteta juurdepääs oma veebilehtedele ja -teenustele. Selle sätte täitmisel on abi kasutajasõbralikest keeletehnoloogiarakendustest, näiteks pimedatele mõeldud kõnesünteesist.

Internetikasutajad ja sisupakkujad saavad keeletehnoloogiast kasu ka vähem ilmsel viisil, näiteks saab seda kasutada veebilehtede automaatselt teise keelde tõlkimisel. Arvestades inimtõlke kõrget hinda, on nõudlusega võrreldes reaalselt kasutatavat keeletehnoloogiat võrdlemisi vähe arendatud ja rakendatud. Selle põhjuseks võib olla eesti keele suhteline keerukus ja tüüpilistes keeletehnoloogiarakendustes kasutatavate tehnoloogiate paljusus.

Järgmises peatükis anname sissejuhatuse keeletehnoloogiasse ja selle põhivaldkondadesse, samuti hinnangu eesti keelt toetava keeletehnoloogia hetkeolukorra kohta.

4

EESTI KEELE KEELETEHNOLOOGILINE TUGI

Keeletehnoloogiaks, sageli kasutatakse ka nimetust "inimkeeletehnoloogia" (ingl k *human language technology*), nimetatakse inimkeele käsitlemiseks loodud tarkvarasüsteeme. Keelel on nii suuline kui ka kirjalik vorm. Kõne on neist vanem ja evolutsiooniliselt loomulikum, samas just kirjalikud tekstid säilitavad keerukat informatsiooni ja enamikku inimeste teadmistest. Kõne- ja tekstitehnoloogiad töötlevad (ja ka genereerivad) keele eri vorme, kasutades selleks sõnastikke, grammatikareegleid ja semantikat. Seega väljendavast meediast (kõne või tekst) sõltumata ühendab keeletehnoloogia keele erinevaid teadmisi. Joonis 1 illustreerib keeletehnoloogia maastikku.

Suheldes kombineerime keelt teiste kommunikatsiooni- ja informatsioonimeediatega, näiteks vestluses kasutame žeste ja miimikat. Digitaalne tekst on ühendatud pildi ja heliga. Film sisaldab nii suulises kui kirjalikus vormis olevat keelt. Teisiti öeldes, kõne- ja tekstitehnoloogiad kattuvad teineteisega ja on omakorda seotud multimodaalset suhtlust ja multimeedia dokumente töötlevate tehnoloogiatega.

Järgnevalt vaatleme peamisi keeletehnoloogia rakenduste valdkondi: keeleline kontroll, veebiotsing, kõnetehnoloogia ja masintõlge. Nad hõlmavad rakendusi ja baastehnoloogiaid, nagu näiteks:

- õigekirjakontroll,
- kirjutaja abivahendid,
- arvutitoetatud keeleõpe,
- infootsing,
- info ekstraheerimine,
- automaatne sisukokkuvõtete tegemine,
- küsimustele vastamine,
- kõnetuvastus,
- kõnesüntees.

Keeletehnoloogia on väljakujunenud uurimisala, millel on märkimisväärne hulk sissejuhatavat kirjandust. Huvitatud lugeja võib tutvuda järgmiste viidetega: [19, 20, 21, 22, 23].

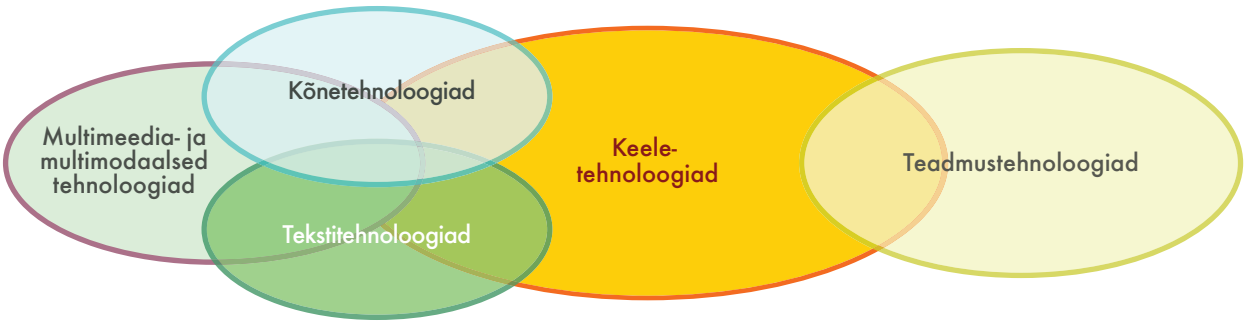
Enne mainitud rakenduste tutvustamist kirjeldame tüüpilise keeletehnoloogilise süsteemi arhitektuuri.

4.1 RAKENDUSTE ARHITEKTUUR

Keeletöötlustarkvara komponendid vastavad keele erinevatele tahkudele. Joonis 2 illustreerib tüüpilise tekstitöötlussüsteemi lihtsustatud arhitektuuri. Kolm esimest moodulit tegelevad tekstisisendi struktuuri ja tähendusega:

1. Eeltöötlus puhastab andmed, analüüsib või eemaldab vorminduse, tuvastab sisendkeele jne.
2. Grammatiline analüüs leiab sõnaliigid, öeldise, sihitise, laiendid, teised lauseliikmed ning tuvastab lause struktuuri.
3. Semantilise analüüsi käigus toimub ühestamine (s.o sõnade konteksti sobivate tähenduste tuvastamine), anafooride lahendamine (nimisõnade vastavusse seadmine asesõnadega), väljendite asendamine ning lause tähenduse esitamine masinloetaval kujul.

Tekstianalüüsi järel alustavad tööd ülesandespetsiifilised moodulid nagu automaatne sisukokkuvõtte tegija ja



1: Keeletehnoloogia infotehnoloogia kontekstis

andmebaasiotsing. See lihtsustatud ja idealiseeritud kirjeldus näitlikustab keeletehnoloogiliste rakenduste arhitektuuri keerukust.

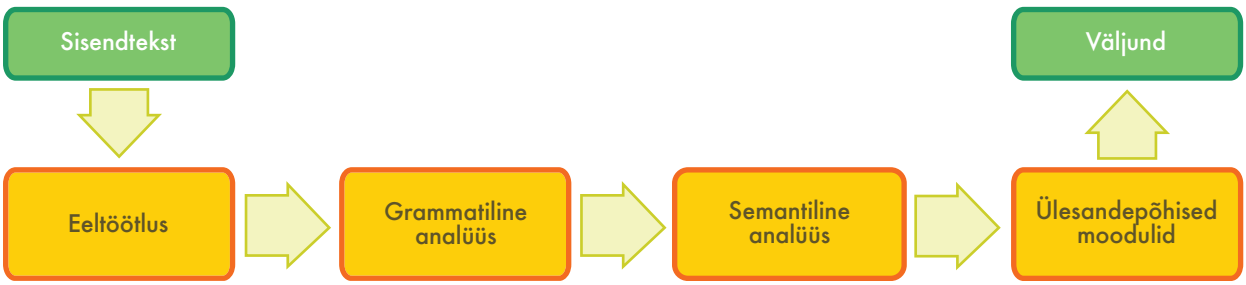
Pärast kesksete keeletehnoloogiliste rakenduste tutvustamist anname ülevaate keeletehnoloogia-alasest uurimistööst ja haridusest ning olnud ja käimasolevatest uurimisprogrammidest. Anname ka eksperthinnangu kesksete rakenduste ja ressursside hetkeseisule erinevates kategooriates, näiteks kättesaadavus, küpsus ja kvaliteet. Tabelis võtame kokku eesti keele keeletehnoloogia üldise hetkeolukorra.

4.2 KESKSED RAKENDUSED

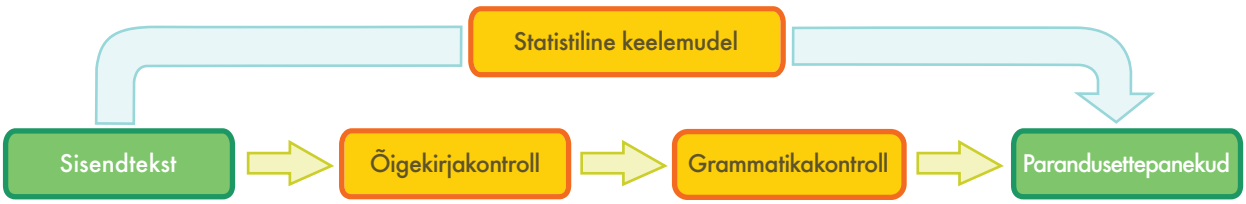
Selles peatükis keskendume kõige olulisemate keeletehnoloogiliste vahendite ja ressursside kirjeldamisele ja anname ülevaate keeletehnoloogia-alasest tegevusest Eestis. Tekstis rõhutatud vahendeid ja ressursse on kirjeldatud ka peatüki lõpus olevas tabelis.

4.2.1 Keeleline kontroll

Igaüks, kes on kasutanud tekstiredaktorit (nt Microsoft Word'i), teab, et sellel on olemas õigekirjakontrollija, mis joonib alla kirjavead ja annab soovitusi nende parandamiseks. Esimesed õigekirjakorrektorid (ehk spellerid) võrdlesid sisestatud sõnu leksikonis olevate korrektsete sõnadega. Tänapäevased spellerid on keerulisemad. Keelespetsiifilisi **grammatikaanalüüsi** algoritme kasutades leitakse morfoloogilised vead (nt mitmuse moodustamine), süntaksivead, näiteks lausest puuduv tegusõna või aluse ja öeldise ühildumise konflikt (nt *nad kirjutas kirja*). Kuid enamik spellereid ei suuda leida vigu sellisest inglisekeelsest tekstist [24] nagu:



2: Tüüpiline keeletöötluse arhitektuur



3: Keeleline kontroll (üleval: statistiline; all: reeglipõhine)

I have a spelling checker,
It came with my PC.
It plane lee marks four my revue
Miss steaks aye can knot sea

(Siin on tegemist sõnademänguga, sõnad on asendatud teiste samasuguse häälduse sõnadega, nii et iga üksiku sõna kirjapilt on korrektne.)

Taoliste vigade tuvastamine vajab kontekstianalüüsi. Sageli juhtub, et hooletu näpulöök klaviatuuril jätab sõnast ära eesti keele mitmusetunnuse *-d*:

värvilise õied
*värvilise**d** õied*

Sellist tüüpi vigade analüüs vajab kas ekspertide poolt käsitsi koostatud **grammatikat** ja seda kasutavat tarkvara või statistilisi keelemudeleid. Viimasel juhul arvutab mudel vastava sõna lauses paiknemise tõenäosuse (st sõna eelneva ja järgneva sõna vahel paiknemise tõenäosuse). Näiteks *värvilise õie* on tunduvalt tõenäolisem sõnade järjend kui *värvilise õied*. Samuti parandab speller otsinguteenuste päringuid, näiteks Google'i *Kas mõtlesite ...*-soovitused.

Automaatselt saab statistilist keelemudelit genereerida siis, kui on olemas suur (korrektsete) tekstide kogum (seda nimetatakse **tekstikorpuseks**). Kirjeldatud meetodeid on kasutatud inglise keele analüüsimiseks. Kahjuks ei ole nad otseselt eesti keelele ülekantavad, sest eesti keelel on vabam sõnajärg, peaaegu piiranguteta liitsõnade moodustamine ning suurem käände- ja pöördelõppude hulk.

Keelelist kontrolli kasutatakse ka mujal kui tekstiredaktorites.

Eesti keele spelleri loomine algas 1991. aastal ning see on olnud tihedalt seotud eesti keele morfoloogiaanalüsaatori ESTMORF arenguga. Spelleri ja morfoloogiaanalüsaatori aluseks on 36000-sõnaline leksikon ja reeglid kõikide sõnavormide moodustamiseks. 1994. aastal anti välja esimene versioon eesti keele spellerist. Hilisemates versioonides on leksikoni täiendatud nimede, lühendite ja neologismidega.

Speller on integreeritud kontoritarkvarapakettidesse MS Office, Open Office.org ja IBM Lotus Notes. Spellerit arendab erafirma Filosoft OÜ [25].

Eesti keele jaoks on püütud luua ka teisi, vabavaralisi spellereid. Tuntuim neist on leksikon ispelli jaoks. Kahjuks ei suuda need spellerid piisavalt edukalt liitsõnu analüüsida.

Grammatikakontrollija kontrollib lause struktuuri ja punktuatsiooni. Eesti keele grammatikakontrollija arendustööga alustati Tartu Ülikoolis 2007. aastal. Hetkel on olemas selle prototüüpversioon, mis suudab tuvastada komavigu 95% täpsusega.

Lisaks tekstiredaktorile kasutatakse keelelist kontrolli ka kirjutaja abivahendites. Need on tarkvarasüsteemid, millega koostatakse etteantud formaadis infotehnoloogia, meditsiini- ja tehnoloogiavaldkondade kasutajajuhendeid ning dokumentatsiooni. Ettevõtted on hakanud oluliselt suuremat tähelepanupöörama nii rahvus-

vahelise turu vajadustele tõlkimise ja lokaliseerimise vallas kui ka tehnilise dokumentatsiooni kvaliteedile. Kehvasti koostatud kasutusjuhendid põhjustavad toodete valesti kasutamist ning sellega kaasnevad klientide kahjunõuded. Keeletehnoloogia arengu käigus on loodud kirjutajaabivahendeid, mis aitavad tehnilise dokumentatsiooni koostajal kasutada piiratud sõnavara ja lausestruktuure, mis vastavad firma kehtestatud nõuetele ja (korporatiiv)terminoloogiale.
Spellerite ja kirjutajaabivahendite kõrval vajab keelelist kontrolli ka arvutitoetatav keeleõpe.

4.2.2 Veebiotsing

Keeletehnoloogia kõige laialtlevinum rakendus on otsing, nii veebis, sisevõrkudes kui ka digitaalsetes raamatukogudes. 1998. aastast tegutsev Google'i otsingumootor teostab praegu umbes 80% kõigist päringutest [26]. 2009. aastast alates on sõna guugeldama lisatud ka Eesti Õigekeelsussõnaraamatusse. Google'i otsinguliidese ja vastuse kuvamise lehekülje kujundus ei ole algusaegadega võrreldes oluliselt muutunud, kuid on toimunud sisulised muutused. Praegune versioon pakub valesti kirjutatud sõnadele õigekirjasoovitusi ning otsingu korrektsust parandab semantiline otsing, mis seisneb päringu konteksti sõnade tähenduste analüüsis [27]. Google'i edulugu tõestab, et suure hulga andmete ja efektiivse indekseerimistehnikaga annab statistiline lähenemine häid tulemusi.

Järgmise põlvkonna otsimootorid peavad kasutama palju keerulisemat keeletehnoloogiat.

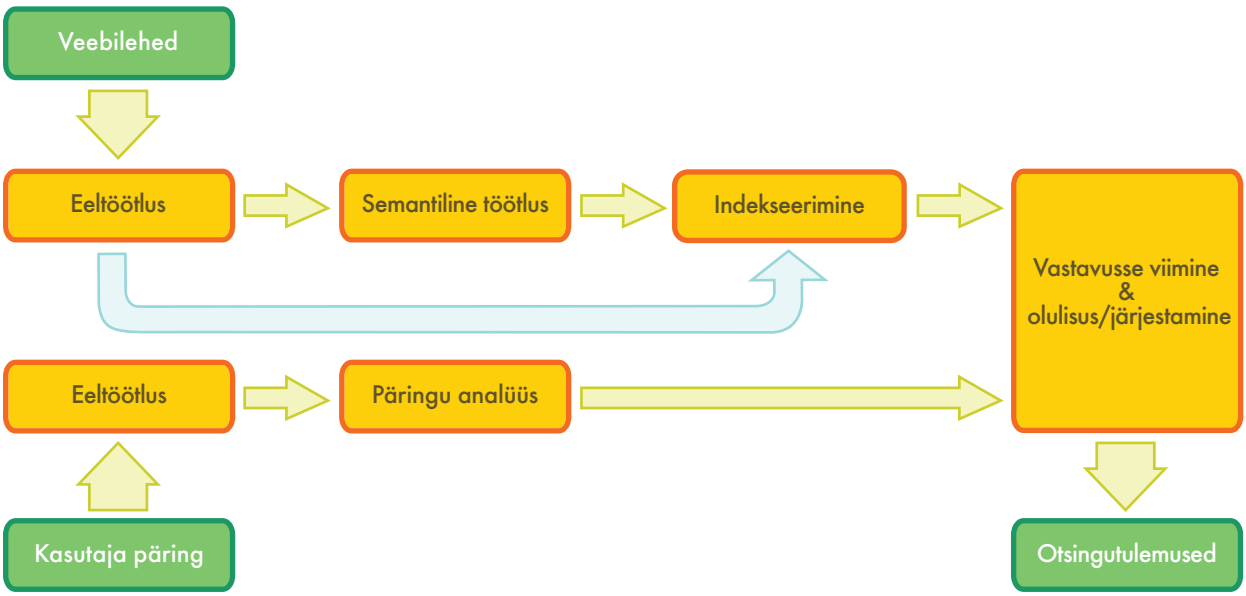
Keerulisema informatsioonivajaduse rahuldamiseks täiendatakse teksti tõlgendamise süsteeme sügavama lingvistilise teabega. Eksperimendid **leksikaalsete ressursside** (masinloetavad tesaurused või ontoloogilised keeleressursid, nt wordnet) kasutamiseks otsingutel on näidanud, et sobivate lehekülgede leidmine paraneb, sest leitakse ka sünonüüme ja nõrgemaid seosetüüpe sisaldavad lehed, näiteks on seotud *aatomienergia* ja *tuumaenergia*.

Võtmesõnade nimekirja asemel küsimustena või muud tüüpi lausetena esitatud päringute töötlemiseks peaksid järgmise põlvkonna otsingumootorid sisaldama palju keerulisemat keeletehnoloogiat. Et vastata päringule "Anna mulle nimekiri kõigist neist ettevõtetest, mille on teised ettevõtted viimase viie aasta jooksul üle võtnud", peab KT süsteem tegema lauses nii süntaktilise kui ka **semantilise analüüsi** ning andma kiiresti vastavate dokumentide indeksi. Vastuse andmiseks tuleb kõigepealt analüüsida lause grammatilist struktuuri ja mõista, et kasutaja tahab just nimekirju ülevõetud ettevõtetest, mitte ettevõtete omandajatest. Rahuldamaks väljendit "viimase viie aasta jooksul", peab süsteem leidma sobiva aastate vahemiku. Seejärel tükk tüki haaval informatsiooni leidmiseks on vaja sobitada päring meeletu hulga struktureerimata andmetega. Kirjeldatud protsessi nimetatakse infootsinguks, see sisaldab nii otsimist kui ka leitud dokumentide järjestamist. Ettevõtete nimekirja genereerimiseks kasutatakse nimeüksuste tuvastamise protsessi, mille käigus tuvastab süsteem dokumentidest ettevõtte nimeks sobiva sõnajärjendi.

Tunduvalt keerulisem on leida päringule vastust teises keeles olevate dokumentide hulgast. Keeltevaheline infootsing eeldab päringu automaatset tõlkimist kõigisse võimalikesse lähtekeeltesse ja hiljem saadud tulemuste tõlkimist sihtkeelde.

Tänapäeval suureneb pidevalt andmete hulk, mis esinevad mingil muul kujul kui kirjalik tekst ja on tekkinud vajadus multimeedia infootsingu teenuse järele, mis otsiks pilte, audiofaile ja videoandmeid. Audio-ja videofailidest otsimiseks teisendab kõnetuvastusmoodul kõne tekstiks või selle foneetiliseks esituseks, mida saab kasutaja päringuga sobitada.

Keelest sõltumatud otsinguvahendid suudavad leida ainult sõnavorme, millel on päringusõnaga täpselt sama



4: Veebiotsing

kuju või mis sisaldavad päringusõna alamsõnena. Kuna eesti keele morfoloogia on rikas ja lisaks lõppudele võib ka sõna tüvi muutuda, siis on edukaks otsinguks ja indekseerimiseks vaja keelespetsiifilisi vahendeid.

Dokumente hoitakse arvutis kui suur tekstilist andmebaasi. Täistekstiotsing jagatakse kaheks alamülesandeks: indekseerimiseks ja otsimiseks. Indekseerimise protsessis analüüsitakse tekste sõna-sõnalt ja luuakse otsisõnade nimekiri ehk indeks. Otsimisfaasis kasutatakse konkreetse päringu töötlemiseks ainult indeksit, mitte kogu teksti. Indekseerija loob kirje iga dokumendist leitud sõna või termini jaoks, kirjesse salvestatakse ka dokumendi viide ja vahel ka selle sõna asukoht dokumendis. Keelespetsiifilised indekseerijad leiavad enne sõnade indeksisse lisamist nende algvormid ehk lemmatiseerivad otsisõnad. Näiteks sõnavormid *käsi*, *käe*, *kätt* esitatakse indeksis ainult tüvisõna ehk lemma *käsi* kirjena. Mõnel juhul leiab lemmatiseerija ühele sõnavormile mitu algvormi, nt *kuue* algvormideks on nii *kuub* kui ka *kuus*. Sellise mitmesuse lahendamiseks otsib süsteem sõnade konteksti põhjal õige algvormi (protsessi nimetatakse morfoloogiliseks ühestamiseks).

Eesti Infosüsteemide Amet on avalikult soovitanud kasutada Eesti avaliku sektori infosüsteemide infootsingul ja indekseerimisel lemmatiseerimismoodulit [2].

Esimene lemmatiseerijat kasutav otsingumootor oli kasutusel 1997–2001 aastal Riigikantselei infosüsteemis. Ka Google'i otsingumootor kasutab eesti keele jaoks mõningast lemmatiseerimist, näiteks päringule *majandusminister* antakse vastuses viiteid ka dokumentidele, milles esineb ainsuse omastavas käändes vorm *majandusministri*.

4.2.3 Suuline suhtlus

Suuline suhtlus on rakendusvaldkond, mis sõltub kõnetehnoloogiast ehk suulise keele töötlemise tehnoloogiast. Suulise suhtluse tehnoloogiat kasutatakse sellise kasutajaliidese loomiseks, kus traditsioonilise graafilise kujunduse, hiire ja klaviatuuri asemel suheldakse arvutiga suulist kõnet kasutades. Tänapäeval kasutatakse näi-

teks hääljuhitavaid kasutajaliideseid osaliselt või täielikult automatiseeritud telefoniteenustes. Hääljuhitavad kasutajaliidesed on kasutusel panganduses, tarneahelate juhtimises, ühistranspordis, telekommunikatsioonis ja teistes ärivaldkondades. Suulise suhtluse tehnoloogiat kasutatakse ka autode navigeerimissüsteemides ning nutitelefonides graafilise puutetundliku kasutajaliidese alternatiivina.

Suulise suhtluse tehnoloogiat kasutatakse sellise kasutajaliidese loomiseks, kus traditsioonilise graafilise kujunduse, hiire ja klaviatuuri asemel suheldakse arvutiga suulist kõnet kasutades.

Suuline suhtlus hõlmab nelja tehnoloogiat:

1. Automaatne **kõnetuvastus** teeb kasutaja poolt kuuldavale toodud helijärjendi põhjal kindlaks tegelikult öeldud sõnad.
2. Loomuliku keele mõistmise protsess analüüsib öeldu süntaktilist struktuuri ja tõlgendab seda vastavalt süsteemi vajadustele.
3. Dialoogi haldamise moodul määrab süsteemi funktsionaalsust arvestades selle, milline tegevus algatatakse vastuseks kasutaja sisendile.
4. **Kõnesüntees** teisendab süsteemi vastuse helideks.

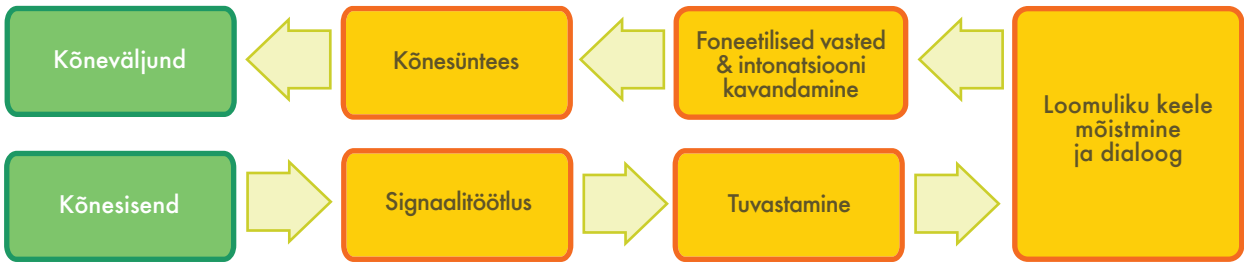
Kõnetuvastussüsteemi suurimaks väljakutseks on kasutaja öeldud sõnade tuvastamine. Probleemi lahendamiseks piiratakse võimalike ütluste hulka konkreetsete võtmesõnadega või siis luuakse käsitsi rohkelt loomuliku keele ütlusi sisaldav keelemudel. Masinõppetehnoloogiaga on võimalik keelemudeleid ka automaatselt luua, selleks kasutatakse **kõnekorpust**, mis koosneb suurest hulgast kõnet sisaldavatest audiofailidest ja teksti transkriptsioonidest. Sõnavara piiramine sunnib inimesi kasutama väga jäika hääljuhitavat kasutajaliidest. Kasutajatele ei pruugi see küll meeldida, kuid samas rikkama sõnavaraga keelemudeli loomine, sobitamine ja ka haldamine on oluliselt kallim. Kasutajatele on vastuvõetavamad keelemudelil põhinevad kasutajaliidesed, mis lubavad neil oma soove võimalikult paindlikult väljendada, näiteks kasutajaliides alustab dialoogi lausega *"Kuidas ma saan sind aidata?"*.

Hääljuhitavate kasutajaliideste tootjad eelistavad väljundi genereerimisel kasutada eelsalvestatud professionaalsete diktorite ütlusi. Staatiliste ütluste korral, mil sõnastus ei sõltu kontekstist ega kasutaja andmetest, annab see parema tulemuse. Dünaamilise sisu korral on tulemus ebaloomuliku intonatsiooniga, sest audiofaili tükid liidetakse lihtsalt kokku. Tänapäeva kõnesünteesisüsteemides on loomulikult kõlavate dünaamiliste ütluste genereerimine muutunud üha paremaks, kuid arenguruumi veel on.

Turul olevate kõnetehnoloogialiideste komponendid on viimase kümnendi jooksul standardiseerunud ning kõnetuvastuse ja kõnesünteesi turg on märkimisväärselt konsolideerunud. G20 riikide rahvuslikel turgudel domineerivad viis globaalset tegijat, Euroopas on neist tuntuimad Nuance (USA) ja Loquendo (Itaalia). 2011. aastal teatas Nuance, et omandas Loquendo, see märgib konsolideerumise jätkumist.

Eesti keele automaatse kõnetuvastusega tegeleb peamiselt Tallinna Tehnikaülikooli Küberneetika Instituudi foneetika ja kõnetehnoloogia labor. 2000. aastal valmis prototüüp isoleeritud sõnade tuvastamiseks (eestikeelsed numbrite ja tähtede nimetused), 2002–2004 valmis piiratud sõnavaraga peidetud Markovi mudelil (HMM) põhinev sidusa kõne tuvastussüsteem. Viimane kõnetuvastussüsteemi versioon (2010) võimaldab tuvastada piiramata sõnavara 63–85% täpsusega. Tulemus sõltub kõne žanrist, sõnavarast ja signaali kvaliteedist (müra tasemest) [28].

On loodud kõnetuvastaja veebirakendus, mis võimaldab automaatselt transkribeeritud raadiovestlussaateid lehitseda, neid kuulata ja nendest otsida. Samuti on olemas veebiteenus, millega kasutaja saab saata süsteemile oma helifaile transkribeerimiseks. Arendamisjär-



5: Kõnepõhine dialoogsüsteem

gus on radioloogidele sobiva kõnetuvastussüsteemi loomine, millega on võimalik dikteerida ka spetsiifilisemat sõnavara. Esialgsed eksperimenditulemused on paljulubavad (10% vigu reaalajalisel tuvastamisel).

Aastatel 1997–2002 loodi kolme organisatsiooni (TTÜ Küberneetika Instituut, Eesti Keele Instituut ja OÜ Filosoft) poolt eesti keele tekst-kõnesüntesaator. See kõnesüntesaatori versioon kuulub n.ö süntesaatorite esimesse põlvkonda, kasutatakse difoone, iga kõneüksus vastab täpselt ühele andmebaasis olevale difoonile (helilt helile üleminekule). Süntesaatori väljund on arusaadav, kuid on monotoonne, veidi hakitud ja pisut ebaloomuliku kõlaga. Süntesaator on kohandatud kasutamiseks pimedatele. Süntesaator on avatud lähtekoodiga, seda võib kasutada mitteärilistel ja mittesõjalistel eesmärkidel [29].

Eesti Keele Instituut arendab hetkel ka korpusepõhise kõnesüntesaatori versiooni, milles lisaks difoonidele kasutatakse ka pikemaid kõneüksusi (sõnu ja fraase).

Haridus- ja teadusministeeriumi parima keeleteo auhinna võitsid 2010. aastal MTÜ Jumalalaegas ja Eesti Hoiuraamatukogu töörühm, kes lõid eestikeelse hääljuhendamise pimedate tehnilistele abivahenditele. Nende rakendused kasutavad soome kõnesüntesaatorit.

Tulevikus on oodata märkimisväärseid muutusi kõnetehnoloogia arengus. Kõnetehnoloogia kasutamist mõjutab ka laialt levima hakanud nutitelefon, mis on tavalise telefoniside, interneti ja e-maili kõrval uus sobiv platvorm kliendisuhete halduseks. Ilmselt on tulevikutelefonis vähem hääljuhitavaid kasutajaliideseid ning suuline kõne hakkab mängima nutitelefonides suuremat rolli kasutajasõbraliku sisendina. Arengu protsess sõltub kõnelejast sõltumatute kõnetuvastussüsteemide korrektsuse paranemisest. Juba praegu pakutakse nutitelefonide kasutajatele tsentraliseeritud teenustena kõne dikteerimist. Sarnased Tanel Alumäe ja Kaarel Kaljuranna TTÜ Küberneetika Instituudis välja töötatud eestikeelsed kõnetuvastusrakendused nutitelefonidele võitsid 2011. aasta parima keeleteo auhinna.

4.2.4 Masintõlge

Mõte kasutada arvuteid loomuliku keele tõlkimiseks tekkis juba 1946. aastal. Olulisel määral rahastati seda uurimissuunda viiekümnendatel ja kaheksakümnendatel aastatel, kuid vaatamata pikale ajaloole ei täida isegi tänapäevane **masintõlge** algselt talle seatud eesmärki, milleks oli automaatne piirideta tõlge.

> Kõige sirgjoonelisem masintõlke viis seisneb ühe keele sõnade asendamises teise keele sõnadega.

Kõige sirgjoonelisem masintõlke viis seisneb ühe keele sõnade asendamises teise keele sõnadega. Selline lähenemine sobib piiratud sõnavaraga valdkondade tekstide (nt ilmateadete) tõlkimiseks. Vähem standardiseeritud

teksti kvaliteetseks tõlkeks on vajalik suuremale tekstiüksusele (fraasile, lausele või tervele lõigule) sobiva sihtkeelse vaste leidmine.

Peamiseks takistuseks on inimkeele mitmesus, mis esitab väljakutse erinevatel analüüsitasanditel, näiteks sõnatähenduse mitmesus leksikaalsel tasandil (*hiir* võib olla nii loom kui arvuti osa) või lause struktuuri mitmesus süntaktilisel tasandil, vt alljärgnevaid tõlkeid inglise keelest:

The woman saw the car and her husband, too.

- *[Naine nägi autot ja tema abikaasa samuti.]*
- *[Naine nägi autot ja samuti oma abikaasat.]*

Masintõlkesüsteem võib põhineda ka lingvistilistel reeglitel. Lähedalt seotud keelte tõlkimisel saab kasutada otsest asendamist. Reeglipõhised (või lingvistiliste teadmiste põhised) masintõlkesüsteemid analüüsivad lähtekeelset teksti ning loovad selle põhjal vahepealse sümbolilise esituse hilisemaks sihtkeelsesse teksti genereerimiseks. Taolised süsteemid vajavad heaks tõlkeks nii põhjalikke leksikone, milles on esitatud morfoloogiline, süntaktiline ja semantiline informatsioon kui ka mahukaid käsitsi koostatud grammatikaid. Vajalike vahendite loomise protsess on pikk ja seetõttu ka kallis.

Hilistel kaheksakümnendatel, kui arvutusvõimsus suurenes ja ühtlasi ka odavnes, tekkis huvi statistiliste masintõlkemudelite loomise vastu. Statistilised mudelid saadakse kakskeelsete tekstikorpuste analüüsil. Näiteks Europarli **paralleelkorpus** sisaldab Euroopa Parlamendi väljaandeid 21 Euroopa keeles. Piisava andmehulga korral leiab masintõlkesüsteem võõrkeelsele tekstile sellise tõlke, mis annab edasi teksti ligikaudse tähenduse. Erinevalt reeglipõhistest süsteemidest genereerib statistiline masintõlkesüsteem sageli grammatiliselt mittekorrektse väljundi. Samas statistilise süsteemi loomiseks on vaja vähem inimtööjõudu ning see katab ka teatud keele eripärasid (nt idiomaatilised väljendid), mida teadmistepõhised süsteemid ignoreerivad.

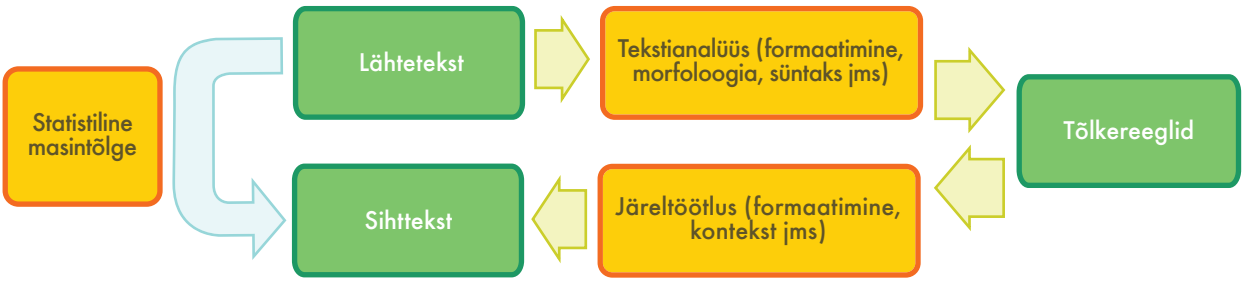
Eesti keele masintõlge on tõsine väljakutse.

Statistiliste ja reeglipõhiste masintõlkesüsteemide tugevad ja nõrgad küljed kompenseerivad üksteist, seetõttu pööratakse hetkel suurt tähelepanu mõlemat lähenemist kombineerivale hübriidsele meetodile. Üheks selle rakendamise võimaluseks on tõlkida paralleelselt lingvistilist ja statistilist tõlget kasutades ja hiljem valikumoodulis otsustada, kumb tõlge on parem. Pikemate lausete (üle 12 sõna) korral on tulemused perfektsusest kaugel. Kvaliteetsema tulemuse saaks kombineerides kummagi tõlke parimaid osi, samas on see küllaltki keeruline ning alati ei ilmne omavahel täpses vastavuses olevad osad.

Eesti keele masintõlge on tõsine väljakutse. Sõnastikupõhise analüüsi muudab keeruliseks vaba liitsõnamoodustus, uusi sõnu saab liitmise teel alati juurde tekitada. Analüüsiprobleeme põhjustavad ka vaba sõnajärg ja mitmeosalised tegusõnad (ühend- ning väljendverbid). Lisaks kõigele muule on piiratud ka paralleelsete tekstide hulk. Vaatamata sellele kuulub Eesti keel nende ligi 50 maailma keele hulka, mida saab arvuti abil tõlkida [30]. Eesti keele masintõlke ajalugu ulatub tagasi 50ndatesse, kui Tartu Ülikooli matemaatikud katsetasid matemaatiliste tekstide tõlkimist vene keelest eesti keelde. Tolleaegne riistvara (arvuti Ural) töötas kiirusega 100 operatsiooni sekundis. Nõrk arvutusvõimsus oligi üks katsete katkestamise põhjustest.

Praegu on eesti keele jaoks olemas kaks masintõlkesüsteemi. Tuntuim neist on Google'i tõlketeenus. Selle kvaliteet ei ole küll alati küllaldane, kuid võimaldab siiski aru saada teksti üldisest teemast ja põhifaktidest.

Teist masintõlkesüsteemi arendab Tartu Ülikooli uurimisgrupp. Nende uurimistöö keskendub hetkel eesti-inglise masintõlkesuunale. Süsteem (http://masintolge.ut.ee) tõlgib piiratud pikkusega lauseid eesti keelest inglise keelde. Masintõlkesüsteem kasutab avatud lähtekoodiga Mosese dekodeerimismooduleid ja seda tree-



6: Masintõlge (vasakul: statistiline; paremal: reeglipõhine)

nitakse erinevatel eesti-inglise paralleelkorpustel, kaasa arvatud JRC-Acquis ning OPUS.

Masintõlkesüsteem suurendab oluliselt töö produktiivsust, eriti siis, kui süsteem on integreeritud töövoogu ning kohandatud kasutajaspetsiifilise terminoloogiaga. Näiteks Siemens kasutab interaktiivset tõlget toetavaid süsteeme ja Volkswagen kasutab keeleportaali, mis tagab ligipääsu sõnaraamatutele, ettevõttespetsiifilisele terminoloogiale, tõlkemälule ja masintõlketoele.

Masintõlkesüsteemide kvaliteedi parandamisel on veel palju arenguruumi. Väljakutseks on keeleressursside kohandamine konkreetse sektori vajadustega ning tehnoloogia integreerimine töövoo protsessidesse, mis juba kasutavad terminoloogiabaasi ja tõlkemälu.

Hindamiskampaaniad võrdlevad masintõlkesüsteemide kvaliteeti, erinevaid lähenemisi ja süsteemide eri keelepaaride olukorda. Euromatrix+ projekti käigus loodud tabelis (vt joonis 7) on koondatud keelepaariti (iiri keelt ei võrreldud) 22 ELi ametliku keele tulemused. Tulemusi on hinnatud BLEU-punktidega, milles parema tõlke skoor on alati kõrgem [31]. Inimtõlkija koguks ülesandest keskmiselt 80 punkti. Parimad punktisummad (tabelis rohelise ja sinise värviga tähistatud) said keeled, millesse on koostööprojektides palju panustatud ning mille jaoks leidub rohkelt paralleeltekste (nt inglise, prantsuse, hollandi, hispaania ja saksa keel). Tabelis on punasega märgitud halvimad tulemused. Nendele keeltele kas ei ole piisavalt tähelepanu pööratud või need keeled erinevad struktuurilt oluliselt teistest keeltest (näiteks ungari, malta, soome keel ja eesti keel).

4.3 MUUD RAKENDUSALAD

Keeletehnoloogiliste rakenduste loomisel tuleb lahendada suur hulk süsteemis paiknevaid alamülesandeid, mida vahel ei ole kasutajaga suhtlemisel isegi näha. Need moodulid on vastavuses arvutilingvistika erinevate alamteemade uurimisobjektidega.

Keeletehnoloogilised rakendused on sageli suuremate tarkvarasüsteemide osad, tõstes nende funktsionaalsust kulisside tagant.

Näiteks on aktiivne uurimisteema küsimustele vastamine, selle jaoks luuakse eraldi märgendatud korpusi ning korraldatakse teaduslikke võistlusi. Küsimustele vastamise kontseptsioon arenes välja võtmesõnadepõhisest otsingust, mille korral väljastab otsingumootor vastuseks sobivad dokumendid. Idee kohaselt esitab kasutaja konkreetse küsimuse, millele süsteem annab ühe vastuse. Näiteks:

Küsimus: Kui vana oli Neil Armstrong sel ajal, kui ta astus Kuu pinnale?

Vastus: 38.

	Sihtkeeled – Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

7: Masintõlge 22 Euroopa Liidu keele vahel – Machine translation between 22 EU-languages [32]

Kuigi ilmselgelt kuulub küsimustele vastamine veebiotsingu valdkonda, on see praegu katusterminiks sellistele uurimisteemadele nagu küsimuste erinevad liigid, nende käsitlemine ja oletatavat vastet sisaldavate dokumentide analüüsimine ja võrdlemine (kas nad annavad vastandlikke vastuseid?) ning konteksti ignoreerimata spetsiifilise informatsiooni (vastuse) tekstist välja filtreerimine.

Eelnev on omakorda seotud informatsiooni ekstraheerimise (IE, kasutatakse ka mõistet info kurnamine) valdkonnaga, mis oli väga populaarne ja oluline 90ndate alguses, mil arvutilingvistikas hakati eelistama statistilisi meetodeid. IE eesmärgiks on dokumentidest väikeste spetsiifiliste infokildude tuvastamine, näiteks tuvastada uudisartiklitest firmade ülevõtmise põhitegijad. Teine tuntud ülesanne seisneb terroriaktide identifitseerimises. Tekstis leiduv info esitatakse tabelina, milles on näidatud akti sooritaja, sihtmärk, aeg, koht ja intsidendi tulemus. IE keskseks teemaks on valdkonnaspetsiifilise vormi täitmine andmetega. IE on üks näide tagaplaanil töötavast tehnoloogiast, mida saab praktikas erinevatesse rakenduskeskkondadesse lõimida.

Automaatne sisukokkuvõtete tegemine ja **teksti genereerimine** on kaks sellist piiripealset ala, mille rakendused toimivad nii iseseisvate programmidena kui ka toetavas rollis n.ö kulisside taga. Automaatse sisukokkuvõtete tegemise käigus leitakse pikast tekstist oluline informatsioon ja esitatakse see lühema tekstina. Seda võimalust pakub ka näiteks tekstiredaktor Microsoft Word. Statistilisi meetodeid kasutatakse teksti oluliste sõnade (sõnad, mis esinevad tekstis väga sageli, kuid ei ole nii sagedased tavalises keelekasutuses) kindlaks tegemiseks ja enim neid olulisi sõnu sisaldavate lausete leidmiseks. Teksti sisukokkuvõttena esitataksegi just need laused. Kuna sisukokkuvõtte tekst koosneb muutmata kujul algse teksti lausetest, siis on kirjeldatud stsenaariumi järgivad programmid pigem tekstist lausete ekstraheerijad,

väljavõtete tegijad. Teiseks võimaluseks on genereerida täiesti uusi lauseid, mida lähtetekstis ei leidu. See lähenemine nõuab aga teksti sügavamat mõistmist ning seetõttu ei ole ta ei piisavalt robustne ega veakindel.
Teksti genereerimise programmid on harva iseseisvad rakendused. Enamasti on nad integreeritud suurematesse tarkvarasüsteemidesse, näiteks haiglate info- süsteemi, mis kogub, salvestab ja töötleb patsientide andmeid. Andmete põhjal aruannete koostamine on üks paljudest sisukokkuvõtte tegemise rakendustest.
Eesti keele jaoks leidub sisukokkuvõtete tegemiseks ainult prototüüpvahend. Eesti keele automaatne sisukokkuvõtja (EstSum) teeb väljavõtted ühe dokumendi piires. Tekstižanr on samuti piiratud – EstSum eeldab, et sisendtekst kuulub uudiste valdkonda.

4.4 HARIDUSPROGRAMMID

Keeletehnoloogia on väga interdistsiplinaarne uurimisala, milles on kombineeritud keeleteaduse, arvutiteaduse, matemaatika, filosoofia, psühholingvistika ja neuroteaduste ühised kogemused.
Keeletehnoloogia-alast haridust annavad Eestis kaks ülikooli: Tartu Ülikool ja Tallinna Tehnikaülikool [33].

- Tartu Ülikooli eesti ja soome-ugri keeleteaduse üliõpilased võivad õppida arvutilingvistika moodulit nii bakalaureuse- kui ka magistriõppes. See moodul sisaldab keeleteooria ja arvutiteaduse alaseid kursuseid, näiteks programmeerimist lingvistidele. Infotehnoloogia eriala bakalaureuse- ja magistriõppe tudengid võivad õppida keeletehnoloogiat eraldi moodulina. Paljud keeletehnoloogia- ja arvutilingvistika-alased kursused on loodud matemaatika-informaatikateaduskonna ja filosoofiateaduskonna koostöös.
- Tallinna Tehnikaülikoolis ei ole üliõpilaste keeletehnoloogia-alane juhendamine nii laialdane. Mõned informatsioonitehnoloogia doktorandid spetsialiseeruvad kõnetehnoloogia õppele, kasutades selleks individuaalseid õppeprogramme.

2009. a-l rajati kaks doktorikooli, milles osalevad keeleteaduse ja keeletehnoloogia doktorandid: informatsiooni- ja kommunikatsioonitehnoloogia doktorikool (keeletehnoloogia doktorantidele) ja keeleteooria, filosoofia ja semiootika doktorikool (arvutilingvistika doktorantidele).

4.5 RIIKLIKUD PROGRAMMID JA ALGATUSED

Keeletehnoloogia-alase uurimistööga alustati Eestis juba 1950ndatel, kui ülikoolidesse ja uurimislaboritesse jõudsid esimesed suurarvutid. 1990ndate alguses muutusid oluliselt senised finantseerimisviisid ja ka uurimisteemad ning järjepidevas uurimistöös tekkis tagasilangus. Tänu rahvusvahelistes projektides (nt Copernicus) osalemisele elasid keeletehnoloogia uurimisrühmad keerulised ajad siiski küllaltki hästi üle [33].
1990ndate lõpul avanesid uued rahastamisvõimalused:

- Eesti Informaatikakeskuse poolt (1998–2000) algatatud programm "Eesti keeletehnoloogia". Selle programmi raames loodi ka 1999. aastal esimene eesti keele keeletehnoloogia arendamise kava.
- Riiklikud programmid "Eesti keel ja kultuurimälu" (1999–2003) ja "Eesti keel ja rahvuslik mälu" (2004–2008) sisaldasid keeletehnoloogia alamprogramme.
- Keeletehnoloogia võtmeisikud olid samuti kaasatud EL 5. raamprogrammi projekti "eVikingsII: Virtuaalse infoühiskonna tehnoloogiate teadus- ja arenduskeskuse asutamine Eestis" (2002–2005).

2005. aastal koostati riikliku programmi "Eesti keele keeletehnoloogiline tugi" (EKKTT) kava. 2006. aastal käivitas haridusministeerium selle programmi viieks

aastaks (2006–2010). Programmi peaeesmärk oli arendada eesti keele keeletehnoloogilist tuge tasemele, mis lubaks eesti keelel moodsas infoühiskonnas vabalt funktsioneerida. EKKTT rahastas keeletehnoloogia-alast teadus- ja arendustööd, kaasa arvatud taaskasutatavate keeleressursside loomist ja keeletehnoloogilise baastarkvara arendamist, ning keeletehnoloogilise infrastruktuuri kaasajastamist. Programmis rahastatud ressursid ja prototüübid on vabaks kasutamiseks [34].

Ühe projektina kasvas EKKTT riiklikust programmist välja Eesti Keeleressursside Keskus (http://www.keeleressursid.ee), mille eesmärgiks on luua infrastruktuur, mis teeb eestikeelsed keeleressursid ja keeletehnoloogilise tarkvara huvilistele kättesaadavaks. 2011. aasta lõpul moodustati Eesti Keeleressursside Keskus konsortsiumina, kuhu kuuluvad kolm partnerit: Tartu Ülikool, Tallinna Tehnikaülikooli Küberneetika Instituut ja Eesti Keele Instituut.

Hetkel on käimas uus riiklik keeletehnoloogiat toetav programm "Eesti keeletehnoloogia (2011–2017)" [35]. Programm eristub eelnenud EKKTT riiklikust programmist selle poolest, et lisaks tarkvaraprototüüpide ja keeleressursside arendamisele pööratakse eriliselt tähelepanu just keeletehnoloogia rakenduste loomisele. Olemasolevad ning programmi käigus loodavad ressurssid ning tarkvara tehakse kättesaadavaks Eesti Keeleressursside Keskuse kaudu.

Haridus- ja teadusministeerium rahastab ka rohkem teadusele orienteeritud keeletehnoloogilisi projekte, pakkudes sihtfinantseerimist ja Eesti Teadusfondi grante. Arvutiteaduse tippkeskuse (2008–2015) töösse on kaasatud samuti arvutilingviste nii Tartu Ülikoolist kui ka Tallinna Tehnikaülikooli Küberneetika Instituudist.

Eesti on osalenud üle-euroopalise keeleressursside ja -tehnoloogia võrgustiku CLARIN (Common Language Resources and Technology Infrastructure, http://www.clarin.eu) tegevuses alates 2008. aastast. 29. veebruarist 2012 on CLARINi organisatsioonivormiks ERIC (European Research Infrastructure Consortium) ning Eesti kui CLARIN ERICu liikme kohustusi hakkab ellu viima Eesti Keeleressursside Keskus, mis kuulub riikliku tähtsusega teaduse infrastruktuuri objektide hulka.

4.6 VAHENDITE JA RESSURSSIDE KÄTTESAADAVUS

Tabel 8 võtab kokku eesti keele keeletehnoloogilise toe hetkeseisu. Oma ala eksperdid hindasid olemasolevaid vahendeid ja ressursse vastavalt seitsmele kriteeriumile skaalal 0 (väga madal) kuni 6 (väga kõrge).
Eesti keeletehnoloogia hetkeseisu analüüsi võib võtta kokku järgnevalt:

- Eesti keele jaoks on olemas nii kõnetuvastuse kui ka -sünteesi vahendid. Nende edasine arendustöö on hetkel aktiivselt käimas. Kõnetuvastuse ja kõnesünteesi vahendid on loodud uurimisasutustes, seetõttu on nad pigem prototüübid kui valmis tooted.
- Vaatamata eesti keele keerulisele morfoloogiale, on eesti keele morfoloogiaanalüsaatori efektiivsus võrreldav teiste Euroopa keelte vastavate vahenditega. Kuna parim morfoloogiaanalüsaator on loodud kommertstarkvarana, siis ei ole see laiemale üldsusele vabalt kasutatav. Teised, vaba tarkvarana loodud analüsaatorid, on tagasihoidlikumate näitajatega ning pole laialdases kasutuses. Eesti keele süntaksianalüsaatorid põhinevad ühel samal reeglipõhisel formalismil, selle baasil loodud grammatika on kohandatud erinevate tekstiliikide analüüsiks. Süntaksianalüsaatoritel on edaspidi veel palju arenguruumi. Semantikat on raskem analüüsida kui süntaksit ning tekstisemantika töötlus on keerulisem kui sõna- ja lausesemantika. Üldiselt on semantilised vahendid ja ressursid saanud madalad hinded. Seega oleks vaja programme ja algatusi, et kiirendada selle ala arengut nii baasuurimistöö kui ka korpuste mahu suurendamise osas.

- Tekstitõlgendamise programmid vajavad mahukat semantilist analüüsi ning eesti keele jaoks on need alles loomisjärgus.
- Keele genereerimise vahenditest on olemas ainult morfoloogilise sünteesi programmid.
- Laiem üldsus kasutab masintõlkeks Google'i tõlketeenust. Tartu Ülikoolis on arendamisel ka eesti-inglise masintõlkesüsteem. Ilmselt oleks suur nõudlus ka eesti-vene-eesti masintõlkele.
- Viimastel kümnenditel on loodud märkimisväärne hulk Eesti keele ressursse (korpused, leksikonid, WordNet), seega olukord on küllaltki hea. Eesti keele üldkorpused on väga mahukad ja kõrge kvaliteediga, kuid süntaktiliselt ja semantiliselt märgendatud korpuste maht on veel väike. Alles hiljuti alustati tööd multimeedia korpustega.
- Sageli ei ole uurimistöö tulemusel valminud kõrgekvaliteediline tarkvara või ressurss piisavalt standardiseeritud või on puudu toetav dokumentatsioon. Samuti ei pruugi selle ressursi või vahendi edaspidine hooldus ja säilitamine olla garanteeritud.

Kokkuvõtvalt näitavad tulemused, et eesti keele keeletehnoloogia baastehnoloogiat ja -ressursse (morfoloogiaanalüsaator, morfoloogiline ühestaja, süntaksianalüsaator, kõnetehnoloogia programmid, üldkorpused, puudepank, leksikaalne andmebaas ja kõnekorpused) puudutav olukord on küllaltki hea. Lisaks on loodud programme ning vajalikke ressursse sisukokkuvõtete automaatseks loomiseks, masintõlkeks ning dialoogsüsteemideks. Kuid need vahendid ja ressursid on pigem lihtsakoelised või piiratud funktsionaalsusega. Näiteks leidub paralleelkorpusi vaid väheste keelepaaride jaoks ning needki on piiratud tekstižanrites.

Mis puutub keerukamatesse valdkondadesse nagu tekstisemantika, keele genereerimine ja märgendatud multimodaalsed ressursid, siis eesti keele jaoks põhivahendid ja -ressursid puuduvad. Uurimistöö kõige komplitseeritumate vahendite ja ressursside loomiseks nagu diskursuse töötlus, dialoogihaldus, semantilised ja diskursuse korpused on juba saanud esimesi tulemusi, kuid need ressursid vajavad täiendamist ning ka vahendite kvaliteedi ulatus on piiratud. Enamik neist vahenditest (v.a morfoloogiaanalüsaator) on loodud uurimisasutustes ja neid võib pidada pigem prototüüpideks, mitte valmis toodeteks. Nende arendamist on toetanud mitmed riiklikud keeletehnoloogia-alased uurimisprogrammid, seetõttu on need vahendid vabaks kasutamiseks.

4.7 KEELTEVAHELINE VÕRDLUS

Praegune keeletehnoloogiline tugi varieerub keeliti märkimisväärselt. Erinevate keelte olukorra võrdlemiseks hinnatakse selles peatükis kahte rakendusvaldkonda (masintõlget ja kõnetöötlust) ning nende baastehnoloogiat (tekstianalüüsi), samuti keeletehnoloogiliste rakenduste loomiseks vajalike ressursside taset. 5-pallisüsteemis hindamistulemuste põhjal jagunesid keeled keeletehnoloogilise toe taseme poolest viie hinnangu vahel:

1. Suurepärane
2. Hea
3. Rahuldav
4. Osaline
5. Nõrk või puuduv

Keeletehnoloogist tuge hinnati järgmiste kriteeriumite põhjal:

Kõnetöötlus: olemasolevate kõnetuvastustehnoloogiate kvaliteet, olemasolevate kõnesünteesitehnoloogiate kvaliteet, valdkondade katvus, kõnekorpuste arv ja maht, kõnetehnoloogiliste rakenduste arv ja kättesaadavus.

Masintõlge: olemasolevate masintõlketehnoloogiate kvaliteet, kaetud keelepaaride arv, lingvistiliste nähtuste ja valdkondade katvus, olemasolevate paralleelkorpuste kvaliteet ja maht, masintõlkerakenduste arv ja varieeruvus.

	Kogus	Kättesaadavus	Kvaliteet	Katvus	Küpsus	Jätkusuutlikkus	Kohandatavus
Keeletehnoloogia: vahendid, tehnoloogiad ja rakendused							
Kõnetuvastus	2	5	2.8	2.8	3	3	3
Kõnesüntees	2	5	2.8	2.8	3	2	3
Grammatiline analüüs	2.5	3.5	3.2	2.8	4	2.5	3.5
Semantiline analüüs	1	1.3	0.9	0.9	1.3	1.3	1.7
Teksti genereerimine	0	0	0	0	0	0	0
Masintõlge	3	3	1.4	2.1	3	4	2
Keeleressursid: ressursid, andmed ja teadmusbaasid							
Tekstikorpused	3	5	2.5	2.1	3	3	2.5
Kõnekorpused	2	5	2.1	2.8	4	4	4
Paralleelkorpused	2	2	2.1	1.4	3	3	2
Leksikaalsed ressursid	3.5	4	3.2	2.8	3.5	3.5	3.5
Grammatikad	2	5	2.8	2.8	3	3	3

8: Eesti keele keeletehnoloogilise toe olukord

Tekstianalüüs: olemasolevate tekstianalüüsitehnoloogiate kvaliteet ja katvus (morfoloogia, süntaks, semnatika), lingvistiliste nähtuste ja valdkondade katvus, kättesaadavate rakenduste arv ja varieeruvus, olemasolevate (märgendatud) tekstikorpuste maht ja kvaliteet, olemasolevate leksikaalsete ressursside (nt WordNet) ja grammatikate kvaliteet ja katvus.

Ressursid: olemasolevate teksti-, kõne- ja paralleelkorpuste kvaliteet ja maht, olemasolevate leksikaalsete ressursside ja grammatikate kvaliteet ja katvus.

Joonised 9 kuni 12 näitavad, et kuigi valitsus on viimastel aastatel suurendanud eesti keele keeletehnoloogia toetamist, kuulub eesti keel võrdluses teiste keeltega siiski neljandatesse ja viiendatesse klastritesse. Eesti ja soome keele tulemused on kohati võrreldavad ning eesti keel paikneb pisut kõrgemal kui teised sama kõneleja-arvuga keeled, nagu läti, leedu ja malta keel. Kõik need keeled aga jäävad oluliselt alla arvuka kasutajaskonnaga ja palju uuritud keeltele nagu näiteks saksa või prantsuse keel. Kuid isegi nende keelte keeletehnoloogised ressursid ei küündi kvaliteedilt ja katvuselt inglise keele tasemeni, mis juhib kõigis keeletehnoloogilistes valdkondades. Ja ka inglise keele ressurssides on puudujääke, mis takistab kõrgekvaliteediliste rakenduste loomist.

Selleks, et luua keerukaid rakendusi, nagu näiteks masintõlget, on vaja ressursse ja tehnoloogiaid, mis kataksid laias ulatuses kõik lingvistilised aspektid ja võimaldaksid sisendteksti sügavat semantilist analüüsi.

4.8 JÄRELDUSED

Selles keeleraportite sarjas tegime olulise katse hinnata 30 Euroopa keele keeletehnoloogist tuge ja koostada nende taset võrdlev kõrgetasemeline analüüs. Kui on üles leitud tühimikud, vajadused ja puudujäägid, on Euroopa keeletehnoloogilisel kogukonnal ja kõigil asjaosalistel võimalik kavandada suuremahulist uurimis- ja arendusprogrammi, mille eesmärgiks on luua tõeliselt mitmekeelne tehnoloogilise toega Euroopa.

Nägime, et Euroopa keelte vahel on tohutud erinevused. Kui mõnede keelte ja rakendusalade jaoks on olemas nii hea kvaliteediga tarkvara kui ka ressursse, siis teistel (enamasti väiksematel) keeltel neid pole. Mitmete keelte jaoks puudub tekstitöötluse baastarkvara ja ka ressursid selle loomiseks. Teiste jaoks on küll põhivahendid olemas, kuid pole võimalik investeerida nende keelte semantilisse töötlusse. Seepärast peame pingutama, et luua kõrgekvaliteediline masintõlge kõigi Euroopa keelte vahel.

Eesti keele keeletehnoloogilise olukorra hinnang annab põhjust ettevaatlikuks optimismiks. Eesti keele keeletehnoloogilist uurimistööd on ka varem rahastatud ning on olemas inimesed, kes uurimis- ja arendustööga tegelevad. Kahjuks esindavad Eesti keeletehnoloogiatööstust ainult mõned üksikud väikeettevõtted.

Eesti keele jaoks on olemas hulk tehnoloogiaid ja ressursse, kuid neid on oluliselt vähem kui inglise keele jaoks. Keeletehnoloogiline tugi on täna kaugel sellest, mis ta peaks olema, et toetada tõeliselt mitmekeelset teadmust, mida ühiskond vajab.

Keeletehnoloogia keerukus ning vajadus suure hulga andmete järele on põhjused, miks on vajalik luua uus infrastruktuur ja organiseerida uurimistöö sidusamalt, et kannustada suuremat koostööd ning teadmiste vahetust.

Uurimis- ning arendustöö rahastamine ei ole sageli järjepidev. Lühiajalised koordineeritud programmid kipuvad vahelduma perioodidega, mil rahastamine puudub või on ebapüsiv. Samuti on puudulik üleüldine EL riikide ja Euroopa Komisjoni programmide koordineerimine.

Saame seetõttu järeldada, et on hädavajalik luua suuremahuline, koordineeritud algatus, mis keskenduks Euroopa keelte keeletehnoloogiliste erinevuste tasandamisele.

META-NETi pikaajaline eesmärk on viia kõrgekvaliteediline keeletehnoloogia kõigi keelteni, et saavutada kultuurilise mitmekesisuse läbi poliitiline ja majanduslik ühtsus. Tehnoloogia aitab lammutada olemasolevad barjäärid ja luua sildu Euroopa keelte vahel. See nõuab, et kõik osapooled, nii poliitikas, uurimistöös kui ka ettevõtetes ja ühiskonnas, ühendaks oma jõupingutused tuleviku heaks.

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

9: Kõnetöötlus: 30 Euroopa keele keeletehnoloogilise toe olukord

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

10: Masintõlge: 30 Euroopa keele keeletehnoloogilise toe olukord

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

11: Tekstianalüüs: 30 Euroopa keele keeletehnoloogilise toe olukord

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

12: Kõne- ja tekstiressursid: 30 Euroopa keele olukord

5

META-NETIST

META-NET on Euroopa Komisjoni rahastatud tippteadmiste võrgustik, millel on 54 liiget 33-st Euroopa riigist [36]. META-NET edendab Mitmekeelse Euroopa Tehnoloogiaühendust **META** (*Multilingual Europe Technology Alliance*), mis kujutab endast keeletehnoloogia-alaste professionaalide ja organisatsioonide üha kasvavat kogukonda.

META-NET edendab Euroopa mitmekeelse infoühiskonna tarbeks tehnoloogilist vundamenti, mis:

- võimaldaks suhelda ja teha koostööd erinevates keeltes;
- tagaks igale eurooplasele sõltumata keelest võrdse ligipääsu informatsioonile ja teadmistele;
- rajaks ja edendaks võrgustunud infotehnoloogiat ja selle kasutusvõimalusi.

Võrgustik toetab Euroopat, mis ühendab endas digitaalset turgu ja inforuumi. See ergutab ja edendab mitmekeelse tehnoloogia loomist kõigi Euroopa keelte jaoks. Need tehnoloogiad toetavad masintõlget, sisutootmist, informatsiooni töötlemist ja teadmuse haldamist laia valdkonna teemade ja rakenduste jaoks. Need tehnoloogiad võimaldavad luua ka intuitiivseid keelepõhiseid kasutajaliideseid erineva tehnika jaoks alates kodumasinatest, autodest ja arvutitest kuni robotiteni välja. Alates käivitamisest 1. veebruaril 2010 on META-NET juhtinud mitmeid tegevusi oma kolmel põhilisel tegevussuunal META-VISION, META-SHARE ja META-RESEARCH.

META-VISION edendab dünaamilist ja mõjukat kogukonda, mida ühendab ühine visioon ja ühine strateegiline uurimisplaan (*strategic research agenda*). Selle tegevussuuna põhirõhk on suunatud sidusa ja ühtse keeletehnoloogia kogukonna loomisele, ühendades killustatud ja eriilmelised huvigrupid üle kogu Euroopa. Valge raamatu sarjas ilmuvad keeleraportid on koostatud 29 keele kohta. Jagatud tehnoloogiline visioon on loodud kolmes valdkondlikus töögrupis: tõlkimine ja lokaliseerimine, meedia- ja informatsiooniteenused, interaktiivsed süsteemid. META tehnoloogianõukogu on loodud arutamaks ja ette valmistamaks strateegilist uurimisplaani, mis põhineb kogu keeletehnoloogia kogukonnaga tihedas omavahelises suhtluses loodud visioonil.

META-SHARE loob hajusa ja avatud vahendi ressursside vahetamiseks ja jagamiseks. Hoidlate partnervõrk (P2P) hakkab sisaldama andmeid keeleressursside, keeletöötlusvahendite ja veebiteenuste kohta, andmed on dokumenteeritud meta-andmetega ja on jagatud standardiseeritud kategooriatesse. Ressurssidele pääseb otseselt ligi ja nad on ühtse vormi järgi otsitavad. Kättesaadavate ressursside hulka kuuluvad nii vabad, avatud lähtekoodiga materjalid kui ka piiratud juurdepääsuga, tasulised kommertstooted.

META-RESEARCH ühendab omavahel seotud tehnoloogiavaldkonnad. Ta otsib võimalusi teiste tehnoloogiate saavutuste kasutamiseks, et seeläbi keeletehnoloogia-alast innovatiivset uurimistööd edendada. See tegevussuund keskendub juhtivatele teadussaavutustele masintõlkes, andmekogumises ja andmestike töötlemises ning korraldab keeleressursside hindamist, koostades selleks vahendite ja meetodite ülevaateid ning korraldades tööpajasid ja koolitusi võrgustiku liikmetele.

office@meta-net.eu – http://www.meta-net.eu

1

EXECUTIVE SUMMARY

During the last 60 years, Europe has become a distinct political and economic structure. Culturally and linguistically it is rich and diverse. However, from Portuguese to Polish and Italian to Icelandic, everyday communication between Europe's citizens, within business and among politicians is inevitably confronted with language barriers. The EU's institutions spend about a billion euros a year on maintaining their policy of multilingualism, i. e., translating texts and interpreting spoken communication. Does this have to be such a burden? Language technology and linguistic research can make a significant contribution to removing the linguistic borders. Combined with intelligent devices and applications, language technology will help Europeans talk and do business together even if they do not speak a common language.

Language technology builds bridges.

The only (unthinkable) alternative to this kind of a multilingual Europe would be to allow a single language to take a dominant position, to replace all other languages. One way to overcome the language barrier is to learn foreign languages. Yet without technological support, mastering the 23 official languages of the member states of the European Union and some 60 other European languages is an insurmountable obstacle for Europe's citizens, economy, political debate, and scientific progress. The solution is to build key enabling technologies. Language technology targeting all forms of written text and spoken discourse can help people to collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems. Language technology solutions will eventually serve as a unique bridge between Europe's languages. An indespensable prerequisite for their development is first to carry out a systematic analysis of the linguistic particularities of all European languages, and the current state of language technology support for them.

Around one million people speak Estonian as their mother tongue. Estonian is the only official language in the Republic of Estonia. Practical language usage in Estonia is regulated by the Language Act and the legislation based thereon. At the same time Estonia is well-known by its e-government and e-Estonia policies. The Estonian language is supported by a long tradition of Estonian education and research.

Estonian does not belong to the family of Indo-European languages. The characteristic features of Estonian include the accent on the first syllable, high frequency of vowels as opposed to consonants, three different lengths of vowels and consonants, the lack of grammatical gender and articles, and a basic vocabulary different from that of the Indo-European languages. Estonian has a rich morphological system. The compounding is relatively free and productive in Estonian and a compound is written as one word-form. Derivation is another productive device for forming new lexical items. Although Estonian has been described as an SVO language, its word order is rather free.

Language technology is a key for the future.

The automated translation and speech processing tools currently available on the market fall short of the envisaged goals. The dominant actors in the field are primarily privately-owned for-profit enterprises based in Northern America. As early as the late 1970s, the EU realised the profound relevance of language technology as a driver of European unity. At the same time, national projects were set up that generated valuable results, but never led to a concerted European effort. Supported by larger research programs in the past, there exists a language technology research scene in Estonia.

Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business. Unfortunately, English language model is not easily transferable, as Estonian has a flexible word order, unlimited compound building and a richer inflection system.

Still, as a product of laborious work, there is a reliable speller for Estonian that is implemented into main office software suites.

The Google search engine has so many users among Estonians that since 2009, the verb *guugeldama* has even had an entry in the Eesti Õigekeelsussõnaraamat (Estonian ortographic and explanatory dictionary). Language independent search tools can only find the word forms which have the same form as the query word or include the query word as a substring. As Estonian is a language with rich morphology, and since in addition to the endings the stem itself may vary, the language specific tools as lemmatisers are needed for searching and indexing. It is officially recommended to use Estonian lemmatiser for searching and indexing of full-text databases in information systems of public sector in Estonia [2].

The two main types of systems 'acquire' language capabilities in a similar manner to humans. Statistical (or 'data-driven') approaches obtain linguistic knowledge from vast collections of concrete example texts. The second approach to language technology is to build rule-based systems. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. Drawing on the insights gained so far, today's hybrid language technology mixing deep processing with statistical methods should be able to bridge the gap between all European languages and beyond.

European research in the area of language technology has already achieved a number of successes. For example, the translation services of the European Union now use the Moses open-source machine translation software, which has been mainly developed in European research projects.

Machine translation is particularly challenging for the Estonian language. The potential for creating arbitrary new words by compounding makes dictionary analysis and dictionary coverage difficult; free word order and split verb constructions pose problems for analysis, also, the amount of available parallel texts is limited. In spite of this, Estonian belongs to one of the languages (currently, around 50) which can be translated by computer. In the long term spoken language applications will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepwise improvements in the accuracy of speaker-independent speech recognition via the speech dictation services already offered as centralised services to smartphone users. Such language recognition applications for smart phones developed at Institute of Cybernetics at TUT won the grand prize of Estonian Language Deed 2011.

Language Technology helps unify Europe.

As this series of white papers shows, there is a dramatic difference between Europe's member states in terms of

both the maturity of the research and in the state of readiness with respect to language solutions Even for large European languages like German or English with well studied language technologies there are still many open research issues, thus for Estonian the amount of further research is even more extensive.

In the case of the Estonian language, we can be cautiously optimistic about the current state of language technology support. For Estonian, a number of technologies and resources exist, but far less than for English. Language technology support today is by far not in a state that is needed for offering the support a true multilingual knowledge society needs.

Tools for speech recognition and speech synthesis have been developed by leading research institutions in Estonian LT. Although the automatic morphological analysis of Estonian is complicated, the efficiency of morphological tools (such as tokeniser, lemmatiser and morphological analyser) is comparable to similar tools for other major European languages. The development of syntactic parsers must continue for achieving better performance. The only available tool for Estonian text generation is morphological synthesiser. Most of the users prefer Google machine translation service. Also, Estonian-English machine translation system is under development in the University of Tartu. There would probably be a high demand for the Russian-Estonian and Estonian-Russian automatic translation service. Most of these tools have been developed by research institutions and can be regarded as prototypes, not as mature products. Unfortunately, the industry consists only of a few SMEs. With regard to resources such as reference corpora, lexicons, wordnets and terminologies, the situation is also reasonably good for Estonian since substantial resources have been built in recent decades.

When it comes to more advanced fields like text semantics, language generation, and annotated multimodal data, Estonian clearly lacks basic tools and resources even if some of these are currently under development.

The research and development was supported by different national programmes of language technology what guarantees the availability of these tools and resources.

This white paper series complements the other strategic actions taken by META-NET (see the appendix for an overview). Up-to-date information such as the current version of the META-NET vision paper [3] or the Strategic Research Agenda (SRA) can be found on the META-NET web site: http://www.meta-net.eu.

META-NET's vision is high-quality language technology for all languages that supports political and economic unity through cultural diversity.

2

LANGUAGES AT RISK: A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

We are witnessing a digital revolution that is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology has helped to automate and facilitate many processes:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint has replaced overhead projector transparencies;
- e-mail allows documents to be sent and received more quickly than using a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- web search engines provide keyword-based access;
- online services like Google Translate produce quick, approximate translations;
- social media platforms such as Facebook, Twitter and Google+ facilitate communication, collaboration, and information sharing.

Although these tools and applications are helpful, they are not yet capable of supporting a fully-sustainable, multilingual European society in which information and goods can flow freely.

2.1 LANGUAGE BORDERS HOLD BACK THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. However, there is a strong likelihood that the revolution in communication technology is bringing together people who speak different languages in new ways. This is putting pressure both on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding and access to shareable knowledge. In the global economic and information space, there is increasing interaction between different languages, speakers and content thanks to new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recently, Google+) is only the tip of the iceberg.

A global economy and information space confronts us with different languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language that we do not understand. According to a recent report from the European Commission, 57% of Internet users in Europe purchase goods and services in non-languages; English is the most common foreign language followed by French, German and Spanish. 55% of users read content in a foreign language while 35% use another language to write e-mails or post comments on the Web [4]. A few years ago, English might have been the lingua franca of the Web – the vast majority of content on the Web was in English – but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded.

Surprisingly, this ubiquitous digital linguistic divide has not gained much public attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

2.2 OUR LANGUAGES AT RISK

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our modern languages?

The wide variety of languages in Europe is one of its richest and most important cultural assets.

Europe's approximately 80 languages are one of our richest and most important cultural assets, and a vital part of this unique social model [5]. While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [6].

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investments in language preservation focussed primarily on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum [7]. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport and energy needs among others.

Language technology targeting all forms of written text and spoken discourse can help people to collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us already today to:

- find information with a search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- follow the spoken directions of a navigation system;
- translate web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core enabling technologies are for each European language.

> Europe needs robust and affordable language technology for all European languages.

To maintain our position in the frontline of global innovation, Europe will need language technology, tailored to all European languages, that is robust and affordable and can be tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image of a text using a suitably powered printing press. Human beings had to do the hard work of looking up, assessing, translating, and summarising knowledge. We had to wait until Edison to record spoken language – and again his technology simply made analogue copies.

Language technology can now simplify and automate the processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management, as well as content production, in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for specialised domains, and often exhibit limited performance. However, there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, edutainment packages, libraries, simula-

tion environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas in which language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology helps overcome the "disability" of linguistic diversity.

Language technology represents a tremendous opportunity for the European Union. It can help to address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. However, citizens need to communicate across the language borders of the European Common Market, and language technology can help overcome this final barrier, while supporting the free and open use of individual languages. Looking even further ahead, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to support their own multilingual communities. Language technology can be seen as a form of "assistive" technology that helps overcome the "disability" of linguistic diversity and makes language communities more accessible to each other. Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: Future intelligent robots with cross-lingual language capabilities have the potential to save lives.

2.5 CHALLENGES FACING LANGUAGE TECHNOLOGY

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required.

The current pace of technological progress is too slow.

Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technological challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

2.6 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to process different tongues, let's look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between their parents, siblings and other family

members. From the age of about two, children produce their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear.

Learning a second language at an older age requires more cognitive effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples.

Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules.

Moving now to language technology, the two main types of systems 'acquire' language capabilities in a similar manner. Statistical (or 'data-driven') approaches obtain linguistic knowledge from vast collections of concrete example texts. While it is sufficient to use text in a single language for training, e. g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then "learns" patterns of how words, short phrases and complete sentences are translated.

This statistical approach usually requires millions of sentences to boost performance quality. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate, all rely on statistical approaches. The great advantage of statistics is that the machine learns quickly in a continuous series of training cycles, even though quality can vary randomly.

The second approach to language technology, and to machine translation in particular, is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than 20 years. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. However, due to the high cost of this work, rule-based language technology has so far only been developed for a few major languages.

The two main types of language technology systems acquire language in a similar manner.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focusses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology, particularly in Europe's economic and information space. Although this technology has made considerable progress in the last few years, there is still huge potential to improve the quality of language technology systems. In the next section, we describe the role of Estonian in European information society and assess the current state of language technology for the Estonian language.

3

THE ESTONIAN LANGUAGE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

Around one million people speak Estonian as their mother tongue. Most of the speakers (922,000) live in the territory of modern Estonia, but approximately 160,000 people speak the language also in Russia, US, Sweden, Canada, Finland and many other countries [8]. According to the census of the year 2000, there are 1,370,052 inhabitants the Republic of Estonia, and of those 167,804 speak Estonian as a second language [9]. Estonian is the only official language in the Republic of Estonia.

Around one million people speak Estonian as their mother tongue.

Varieties of Estonian include the regional varieties (dialects and the corresponding written forms); varieties of Estonian spoken by Estonians living abroad, social varieties (various sociolects) and varieties used by people with specific linguistic needs, including sign language.

The biggest regional linguistic differences occur between the Northern and Southern dialects. These differences date from the times before Christ when different languages started to secede from the original Balto-Finnic branch of the Uralic languages. The fact that people were very sedentary until the 19th century conduced to the development of regional language forms which can be divided up to 100 parish dialects. Contemporary Standard Estonian evolved on the basis of the Northern dialects but acquired some features also from the Southern dialects [37].

Nowadays, the regional varieties are mostly spoken in Southern Estonian and on the western islands. The Setu and Võru regional varieties spoken in the South-Eastern corner of Estonia are the most different from Standard Estonian. The state supports the use of regional varieties and their preservation as a cultural treasure, as a development source of Standard Estonian and as bearer of the local identity. Many schools in Võru and Viljandi counties in South Estonia teach local dialects (Võru, Setu and Mulgi) as an optional subject.

The Estonian spoken abroad consists of several varieties all influenced to some extent by the dominating language spoken in the country of residence of the speakers. In some cases the Estonian spoken by an emigrant group abroad has developed on its own for more than a century.

Estonian sign language – more precisely, Estonian sign language and signed Estonian are the mother tongue of nearly 2000 deaf people [11]. It is also one the main means of communication for the people with impaired hearing (and for the attendants of both groups).

3.2 PARTICULARITIES OF THE ESTONIAN LANGUAGE

Estonian belongs to the Finnic branch of the Finno-Ugric languages, along with Finnish, Karelian, and other

closely related languages. Also, Estonian is distantly related to Hungarian. It is important to note that Uralic languages do not belong to the family of Indo-European languages.

Estonian does not belong to the family of Indo-European languages.

Typologically, Estonian represents a transitional form from an agglutinating language to a fusional language. Over the course of Estonian history, German has exercised a strong influence on Estonian, both in vocabulary and syntax. The characteristic features of Estonian include the accent on the first syllable, high frequency of vowels as opposed to consonants, three different lengths of vowels and consonants (so-called gradation), the lack of grammatical gender (even in pronouns) and articles, and a basic vocabulary different from that of the Indo-European languages.

Estonian has a rich morphological system: the nouns decline in 14 cases both in singular and plural; the verbs are inflected for tense, person, mood and voice. Although Estonian has 14 cases, there is no accusative, the typical case for coding the syntactic object, which is instead coded using the partitive, genitive or nominative case forms. The compounding is relatively free and productive in Estonian and a compound is written as one word-form. The compounds are produced on the fly, if needed, and as so there is no possibility of listing all of them in a lexicon. Derivation is another productive device for forming new lexical items.

In Estonian, there is no grammatical future tense and the future actions are most often expressed using a verb in a present tense form with the time being indicated by the context:

Ta saabub homme.
(He/she will arrive tomorrow.)

The conditional and oblique moods of verbs are also quite exceptional typologically. Conditional mood is marked by affix *-ks(i)-* and it expresses hypothetical state of affairs or an uncertain event:

Kui ta treeniks rohkem, jookseks ta kiiremini.
(If he/she trained more, he/she would run faster.)

Verbs in oblique mood are marked with an ending *-vat*. This mood is used to report a non-witnessed event without confirming it:

Ta jooksvat kiiresti.
(It has been told that he/she runs fast.)

Although Estonian has been described as an SVO language, its word order is rather free. The only dominating tendency is to use the finite verb form in the second position in the clause. The word order is affected by the information structure of the clause, i. e. it distinguishes given and new information:

- *Ta jooksis kiiresti koju.*
  (He/she ran home fast.)
- *Kiiresti jooksis ta koju.*
  (It was very fast how he/she ran home.)
- *Koju jooksis ta kiiresti.*
  (He/she ran home fast [but not so fast to school].)
- *Jooksis ta kiiresti koju?*
  (Did he/she run home fast?)
- *Kui ta kiiresti koju jooksis, siis ...*
  (While he/she was running home fast, ...)

Although Estonian is close to Finnish, the long-time German influence has induced several features characteristic to the so-called Standard Average European (SAE) [12]. The SAE-like features include the word order in some clause types, the massive use of particle verbs, especially for coding the aspect of the action. So for saying *He did it* one needs free words in Finnish: *Hän teki sen* and the same words plus an extra aspectual

particle in Estonian: *Ta tegi selle ära*. Estonian vocabulary also contains more foreign words and loanwords than Finnish.

Estonian uses phonetic spelling and is written in Latin alphabet. In addition to the basic character set the extra characters õ, ä, ö, ü are used, also š and ž in foreign words. There is a thorough typological description of the language edited by Mati Erelt [14].

3.3 RECENT DEVELOPMENTS

Estonian has been influenced by German (initially Middle Low German, later also standard German), Russian and English, though it is not related to them genetically. After the World War II, Estonian was a subject to russification. Estonian, which had been the official language since the establishment of the independent state of Estonia in 1918, was downgraded. After the collapse of the Soviet Union, Estonian once again became the only official language of the Republic of Estonia.

Today, Estonian (as also, e. g., Icelandic) is one of the smallest languages in the world functioning as an official language at all levels of social interaction: administration, media, literature, theatre, business, school, universities, research, etc.

During the last decades, after Estonia had regained independence from the Soviet Union, on one hand, the position of Estonian has strengthened: it has the status of official language and its preservation is guaranteed by laws, and on the other hand, the position has been weakened due to the globalisation and the development of the information society. The problems known to many other languages are also seen as the greatest threats to Estonian: the drop in the number of mother-tongue speakers, blurring of the language norms, excessive influence of foreign languages, especially that of the English-speaking and -writing social media and pop-culture. Keeping up with the bigger languages in terms of language technology is also a complicated task.

The problems known to many other languages are also seen as the greatest threats to Estonian: the drop in the number of mother-tongue speakers, blurring of the language norms, excessive influence of foreign languages, especially that of the English-speaking and -writing social media and pop-culture.

In order to preserve Estonian language, several state institutions have been established. The Language Inspectorate checks the enforcement of the legislative acts concerned with the language matters. The Language Policy Department at the Ministry of Education and Research is involved with Estonia's language policy planning and helps to make the language better known abroad. The Estonian Language Council is ministry's advisory council on language and has compiled the strategy for maintaining and developing the Estonian language.

3.4 LANGUAGE CULTIVATION IN ESTONIA

According to the Constitution, the official language of the Republic of Estonia is Estonian and it is the state's obligation to preserve the Estonian language, nation and culture "through the ages". The measures necessary for the preservation and development of the Estonian language are defined in the Development Strategy of the Estonian Language (2004–2010) [37] and the upcoming Estonian Language Development Plan (2011–2017) [15].

Practical language usage in Estonia is regulated by the Language Act and the legislation based thereon.

Practical language usage in Estonia is regulated by the Language Act and the legislation based thereon.

Practical language usage in Estonia is regulated by the Language Act and the legislation based thereon. The ac-

tivities related to the development and usage of Estonian (as well as other languages) is coordinated by the Ministry of Education and Research. The Estonian Language Council observes and analyses the language situation and prepares monitoring reports and follow-up strategies concerning the language strategy. Among the divisions of the Ministry of Education and Research, the National Examinations and Qualifications Centre and the Language Inspectorate are engaged in language-related issues. Among the agencies under the administration of the ministry, such issues are dealt with by the Institute of Estonian Language. The Language Commission of the Mother Tongue Society, the Tartu Language Cultivation Centre and also the professors and researchers of the Tartu and Tallinn Universities are involved in language cultivation.

Estonian is one of the official languages of the European Union and Estonian EU legislative terminology is being developed in cooperation with the terminology department of the Institute of Estonian Language and the Estonian Terminology Society.

In 2003, the Development Strategy of the Estonian Language 2004–2010 was compiled by the members of the Estonian Language Council [37]. The strategy provides a research-based description of the situation of the Estonian language, the objectives that need to be achieved, the necessary steps and institutions and people involved. The development strategy of the Estonian language covers all the major areas of language use including language technology.

The next development plan for Estonian language has been compiled in 2010 by Estonian Language Council [15]. The development plan for Estonian language 2011–2017 is a document which lays down main strategic directions for the development, teaching, researching and protection of Estonian. Together with its application plan, relevant legislative documents and other supporting activities (e. g., financing), the development plan for Estonian language will ensure the status of Estonian as an offcial language and its continuous use as the primary language of communication in the Estonian state.

3.5 LANGUAGE IN EDUCATION

Education is one of the most important means of guaranteeing the development and position of a language. One of the roles of education is to provide general and professional literacy and to shape favorable attitudes towards Estonian language among the non-Estonians. General education, especially compulsory general education, is of fundamental importance because it influences the language use most of all.

According to the law, any language can serve as the tuition language in basic school. At present time, two languages of tuition are used in secondary schools: three quarters of schools teach in Estonian, one quarter in Russian. In order to improve the proficiency of Estonian among non-Estonian secondary school graduates, a transition process for non-Estonian secondary schools has begun in 2007 for teaching some of the topics in Estonian.

Estonian language is a compulsory subject in all basic and secondary schools (including all schools of similar levels which teach various professionals). In fall and spring semester of 2009/2010, the number of students in Estonian basic schools was 90,837 (ca 84,000 being native Estonians), while in secondary schools the number of students was 23,769 (22,741 being native Estonians) [15].

The Estonian language is supported by a long tradition of Estonian higher education and research.

The Estonian language is supported by a long tradition of Estonian higher education and research.

ever, internationalisation has increased the proportion of teaching in foreign languages, and also the number of foreign students and university lecturers. In Estonian universities, almost all fields can be studied in Estonian at all levels. At the bachelor level, the student will almost always study his/her field in Estonian, although some specific topics might be taught in other languages.

Among non-Estonian adults, Estonian language courses have been mostly organised for professionals with the greatest communicative needs (e. g., medical nurses, police officers) and for people applying for Estonian citizenship (tuition fees are refunded to successful learners); TV courses of Estonian are also being organised.

3.6 INTERNATIONAL ASPECTS

Estonian has been one of the official languages of the European Union since 2004. This means that Estonian can be used as a language in international communication.

Estonia is becoming more and more attractive to tourists. As a result, the number of people interested in Estonian language and culture has increased during the recent years.

Estonia supports teaching of Estonian abroad – at the moment there are more than 30 universities offering various opportunities to study Estonian [16].

3.7 ESTONIAN ON THE INTERNET

According to statistics, in 2010 about 381,300 families in Estonia have Internet connection at home, and 758,100 people use Internet regularly [17].

Estonia is well-known by its e-government and e-Estonia policies.

Estonia is well-known by its e-government and e-Estonia policies. e-Governance comes in two parts; the act of governing using the Internet (voting, participation) and the delivery of public services. For example, by using these e-services, citizens/residents of Estonia can vote at elections, declare taxes, register to medical consultations, or even monitor the achievements of their children at school.

Most of the local companies have their web pages in Estonian, the newspapers publish their news in the news portals (e. g., http://postimees.ee, http://ohtuleht.ee, http://paevaleht.ee, and many others) [18]. There are a lot of specialised Internet forums, where users communicate in Estonian. Also the social networks like Orkut and Facebook have been localised to Estonian. In addition, there exist numerous chatrooms where the users communicate in Estonian slang. The voluntary community is developing Estonian Wikipedia with more than 88,900 pages.

The growing importance of the Internet is critical for language technology. The vast amount of digital language data is a key resource for analysing the usage of natural language, in particular, for collecting statistical information about patterns. And the Internet offers a wide range of application areas for language technology.

The most commonly used web application is search, which involves the automatic processing of language on multiple levels as will be shown in more detail later. Web search involves sophisticated language technology that differs for each language.

It is an expressed political aim in Estonia as well as other European countries to ensure equal opportunities for everyone. Public agencies need to make sure that their web sites and Internet services can be used by the disabled without restrictions. User-friendly language technology tools offer the principal solution to satisfy this regulation, for example by offering speech synthesis for the blind.

Internet users and providers of web content can also use language technology in less obvious ways, for example, by automatically translating web page contents from one language into another. Despite the high cost of manually translating this content, comparatively little language technology has been developed and applied to the issue of website translation in light of the supposed need. This may be due to the complexity of the Estonian language and to the range of different technologies involved in typical applications.

The next chapter gives an introduction to language technology and its core application areas, together with an evaluation of current language technology support for Estonian.

4

LANGUAGE TECHNOLOGY SUPPORT FOR ESTONIAN

Language technology is used to develop software systems designed to handle human language and are therefore often called "human language technology". Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 1 illustrates the LT landscape.

When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies.

In this section, we will discuss the main application areas of language technology, i. e., language checking, web search, speech interaction, and machine translation. These applications and basic technologies include

- spelling correction
- authoring support
- computer-assisted language learning
- information retrieval
- information extraction
- text summarisation
- question answering
- speech recognition
- speech synthesis

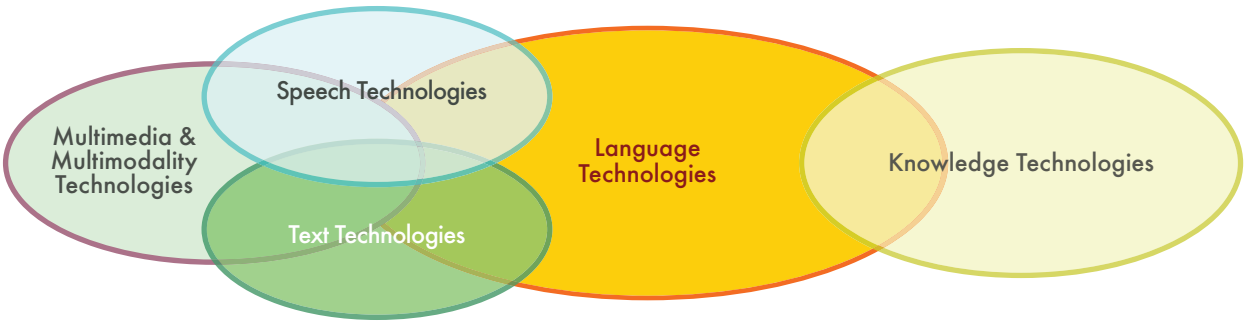
Language technology is an established area of research with an extensive set of introductory literature. The interested reader is referred to the following references: [19, 20, 21, 22, 23].

Before discussing the above application areas, we will briefly describe the architecture of a typical LT system.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. While such applications tend to be very complex, figure 2 shows a highly simplified architecture of a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input languages, and so on.
2. Grammatical analysis: finds the verb, its objects, modifiers and other sentence elements; detects the sentence structure.



1: Language technology in context

3. Semantic analysis: performs disambiguation (i. e., computes the appropriate meaning of words in a given context); resolves anaphora (i. e., which pronouns refer to which nouns); represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups.

In the remainder of this section, we firstly introduce the core application areas for language technology, and follow this with a brief overview of the state of LT research and education today, and a description of past and present research programmes. Finally, we present an expert estimate of core LT tools and resources for Estonian in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Estonian language is summarised in a matrix (figure 7). Tools and resources that are boldfaced in the text can also be found in figure 7 (p. 60) at the end of this chapter. LT support for Estonian is also compared to other languages that are part of this series.

4.2 CORE APPLICATION AREAS

In this section, we focus on the most important LT tools and resources, and provide an overview of LT activities in Estonia.

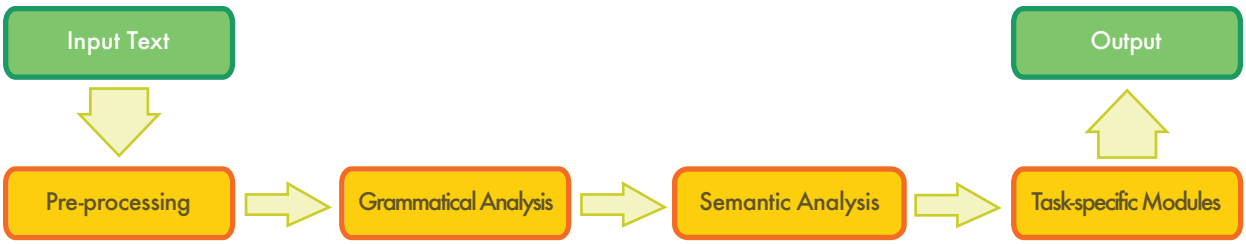
4.2.1 Language Checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spell checker that highlights spelling mistakes and proposes corrections. The first spelling correction programs compared a list of extracted words against a dictionary of correctly spelled words. Today these programs are far more sophisticated. Using language-dependent algorithms for **grammatical analysis**, they detect errors related to morphology (e. g., plural formation) as well as syntax-related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spell checkers will not find any errors in the following text [24]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Handling these kinds of errors usually requires an analysis of the context. For example: deciding whether the noun phrase has an agreement in case and number, as in:

- *värvilise õied*
  colourful-SG-GEN blooms
- *värvilised õied*
  colourful-PL-NOM blooms



2: A typical text processing architecture

This type of analysis either needs to draw on language-specific **grammars** laboriously coded into the software by experts, or on a statistical language model. In this case, a model calculates the probability of a particular word as it occurs in a specific position (e. g., between the words that precede and follow it). For example: *värvilised õied* is a much more probable word sequence than *värvilise õied*. Language checking applications also automatically correct search engine queries, as found in Google's *Did you mean...* suggestions.

A statistical language model can be automatically created by using a large amount of (correct) language data (called a **text corpus**). Most of these two approaches have been developed around data from English. Neither approach can transfer easily to Estonian because the language has a flexible word order, unlimited compound building and a richer inflection system.

The use of language checking is not limited to word processors. It also applies to authoring support systems.

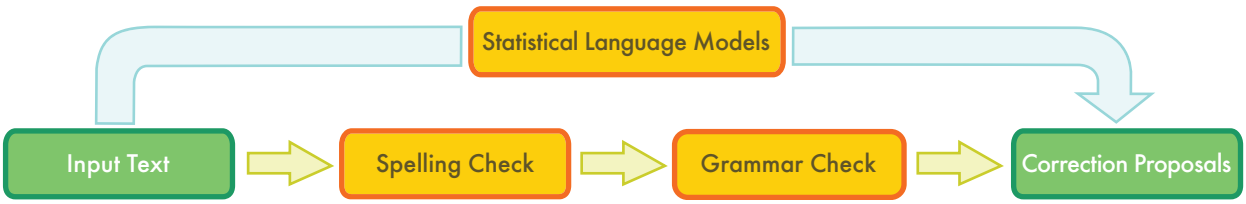
The development of Estonian speller dates back to 1991. The implementation of speller software has been strongly related to the progress of development of morphological analyser for Estonian, known under the name ESTMORF. The base for the speller and analyser was the lexicon of 36,000 simple words and the descriptions for forming all word forms of them. The further development focused on the modeling of compounding and derivation. In 1994, the first version of Estonian speller was released. The later versions have had better lexicons of proper names, abbreviations, neologisms etc.

The speller has been integrated with MS Office, OpenOffice.org and IBM Lotus Notes programs. The speller is being developed by the private company Filosoft [25].

Also, there have been other attempts for developing spellers for Estonian. Several versions of open source tools for spell-checking Estonian have been created. The lexicon for *ispell* is most widely known. Its main disadvantage is the inability to handle compounds.

Grammar checker verifies the structure and punctuation of the sentence. The development of Estonian grammar checker started in 2007 in the University of Tartu. At present time, a prototype of the checker has been created which is able to find places of missing commas with the precision of 95%.

Language checking is not limited to word processors; it is also used in "authoring support systems", i. e., software environments in which manuals and other types of technical documentation for complex IT, healthcare, engineering and other products, are written. To offset customer complaints about incorrect use and damage claims resulting from poorly understood instructions, companies are increasingly focusing on the quality of technical documentation while targeting the international market (via translation or localisation) at



3: Language checking (top: statistical; bottom: rule-based)

the same time. Advances in natural language processing have led to the development of authoring support software, which helps the writer of technical documentation to use vocabulary and sentence structures that are consistent with industry rules and (corporate) terminology restrictions.

Besides spell checkers and authoring support, language checking is also important in the field of computer-assisted language learning.

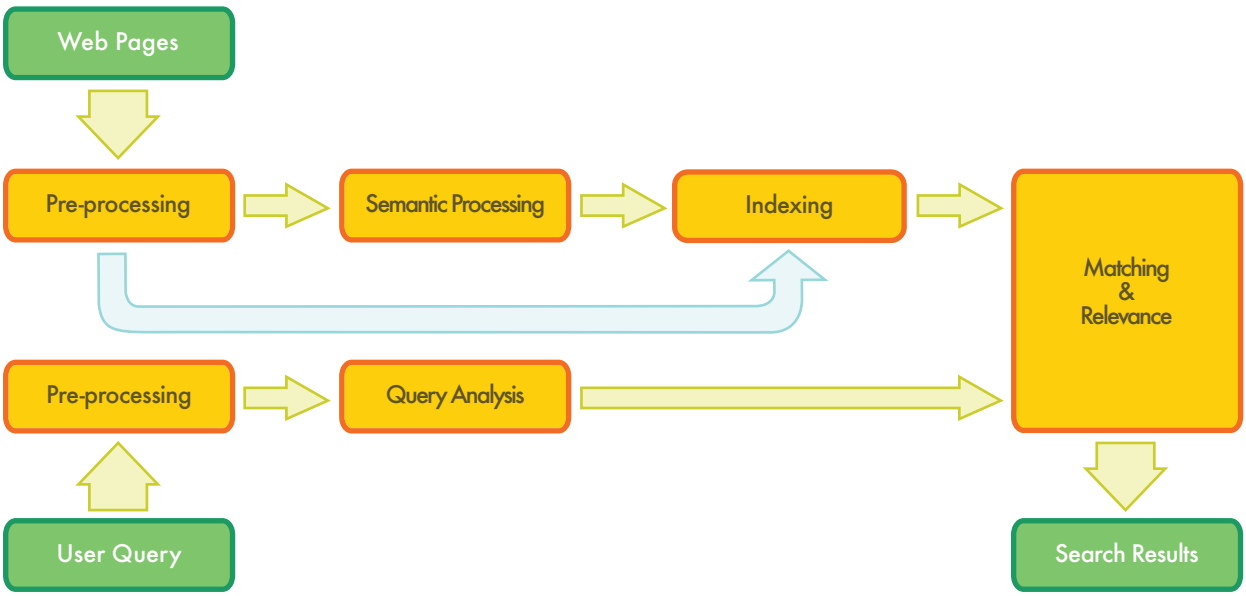
4.2.2 Web Search

Searching the Web, intranets or digital libraries is probably the most widely used yet largely underdeveloped language technology application today. The Google search engine, which started in 1998, now handles about 80% of all search queries [26]. Since 2009, the verb *guugeldama* has even had an entry in the Eesti Õigekeelsussõnaraamat (Estonian ortographic and explanatory dictionary). The Google search interface and results page display has not significantly changed since the first version. However, in the current version, Google offers spelling correction for misspelled words and incorporates basic semantic search capabilities that can improve search accuracy by analysing the meaning of terms in a search query context [27]. The Google success story shows that a large volume of data and efficient indexing techniques can deliver satisfactory results using a statistical approach to language processing.

For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge to facilitate text interpretation. Experiments using **lexical resources** such as machine-readable thesauri or ontological language resources (e. g., WordNet for English) have demonstrated improvements in finding pages using synonyms of the original search terms, such as *aatomienergia* [atomic energy] and *tuumaenergia* [nuclear energy], or even more loosely related terms.

> The next generation of search engines will have to include much more sophisticated language technology.

The next generation of search engines will have to include much more sophisticated language technology, especially to deal with search queries consisting of a question or other sentence type rather than a list of keywords. For the query, *Give me a list of all companies that were taken over by other companies in the last five years*, a syntactic as well as **semantic analysis** is required. The system also needs to provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, rather than companies that have acquired other companies. For the expression *last five years*, the system needs to determine the relevant range of years, taking into account the present year. The query then needs to be matched against a huge amount of unstructured data to find the pieces of information that are relevant to the user's request. This pro-



4: Web search

cess is called information retrieval, and involves searching and ranking relevant documents. To generate a list of companies, the system also needs to recognise a particular string of words in a document represents a company name, using a process called named entity recognition.

A more demanding challenge is matching a query in one language with documents in another language. Cross-lingual information retrieval involves automatically translating the query into all possible source languages and then translating the results back into the user's target language.

Now that data is increasingly found in non-textual formats, there is a need for services that deliver multimedia information retrieval by searching images, audio files and video data. In the case of audio and video files, a speech recognition module must convert the speech content into text (or into a phonetic representation) that can then be matched against a user query.

Language independent search tools can only find the word forms which have the same form as the query word or include the query word as a substring. As Estonian is a language with rich morphology, and since in addition to the endings the stem itself may vary, the language specific tools are needed for searching and indexing.

Documents are held in the computer as one big textual database. The problem of full text search is often divided into two sub-tasks: indexing and searching. The indexing stage will scan the texts of all documents and build a list of search terms, often called the index. In the search stage, when performing a specific query, only the index is referenced rather than the text of the original documents. The indexer will make an entry in the index for each term or word found in a document and possibly its relative position within the document. Language-specific indexers also employ language-specific stemming (lemmatising) on the words being indexed, so for example any token of the Estonian words *käsi*, *käe*, or *kätt* (*hand* in nominative, genitive and partitive case) will be recorded in the index under a single stem-form (lemma) *käsi*. In some cases, automatic lemmatiser may find many lemmas for one word form. For example, the lemmas of *kuue* are *kuub* (jacket) or *kuus* (six). In order

to solve this ambiguity, the system has to take account the context of the word and perform morphological disambiguation process.

It is officially recommended to use Estonian lemmatiser for searching and indexing of full-text databases in information systems of public sector in Estonia [2]. The first use of lemmatiser-based search was implemented 1997–2001 for information systems of State Chancellery. Also Google search for Estonian includes some probably stochastic lemmatisation, e. g., search word *majandusminister* would also return hits for *majandusministri*, the singular genitive of *majandusminister* (minister of economy).

4.2.3 Speech Interaction

Speech interaction is one of many application areas that depend on speech technology, i. e., technologies for processing spoken language. Speech interaction technology is used to create interfaces that enable users to interact in spoken language instead of using a graphical display, keyboard and mouse. Today, these voice user interfaces (VUI) are used for partially or fully automated telephone services provided by companies to customers, employees or partners. Business domains that rely heavily on VUIs include banking, supply chain, public transportation, and telecommunications. Other uses of speech interaction technology include interfaces to car navigation systems and the use of spoken language as an alternative to the graphical or touchscreen interfaces in smartphones.

Speech interaction technology comprises four technologies:

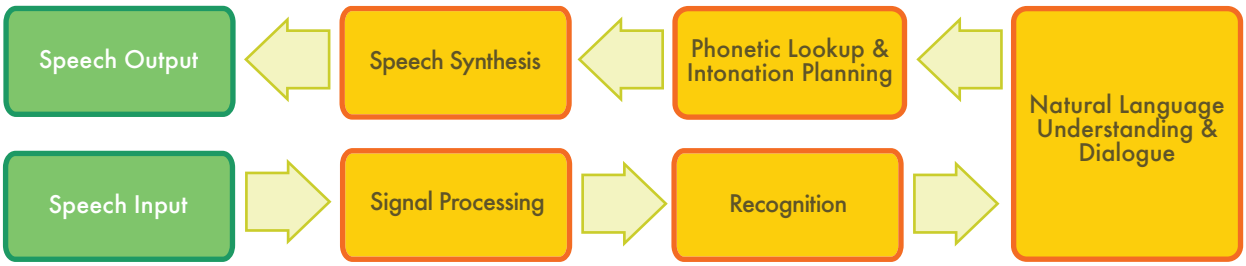
1. Automatic **speech recognition** (ASR) determines which words are actually spoken in a given sequence of sounds uttered by a user.
2. Natural language understanding analyses the syntactic structure of a user's utterance and interprets it according to the system in question.
3. Dialogue management determines which action to take given the user input and system functionality.
4. **Speech synthesis** (text-to-speech or TTS) transforms the system's reply into sounds for the user.

One of the major challenges of ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from **speech corpora**, i. e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly – prompted by a *How may I help you?* greeting – tend to be automated and are better accepted by users.

Speech interaction is the basis for creating interfaces that allow a user to interact with spoken language instead of a graphical display, keyboard and mouse.

Companies tend to use utterances pre-recorded by professional speakers for generating the output of the voice user interface. For static utterances where the wording does not depend on particular contexts of use or personal user data, this can deliver a rich user experience. But more dynamic content in an utterance may suffer from unnatural intonation because different parts of audio files have simply been strung together. Through optimisation, today's TTS systems are getting better at producing natural-sounding dynamic utterances.

Interfaces in speech interaction have been considerably standardised during the last decade in terms of their



5: Speech-based dialogue system

various technological components. There has also been strong market consolidation in speech recognition and speech synthesis. The national markets in the G20 countries (economically resilient countries with high populations) have been dominated by just five global players, with Nuance (USA) and Loquendo (Italy) being the most prominent players in Europe. In 2011, Nuance announced the acquisition of Loquendo, which represents a further step in market consolidation.

During the last decade, the research on automatic speech recognition in Estonia has been carried out mainly at the Laboratory of Phonetics and Speech Technology, Institute of Cybernetics at Tallinn University of Technology (IOC). In 2000, a prototype for isolated word recognition (Estonian numbers and names of Estonian letters) was developed. In 2002–2004 a limited vocabulary connected speech recognition system based on hidden Markov models (HMM) as context-dependent phone units was developed. The last version (2010) of the system with unlimited vocabulary is able to recognize 63–85% of words. The result depends heavily on the genre of the speech, vocabulary and the quality of the signal (the noise level) [28].

There is a web application of speech recogniser which enables to browse automatically transcribed conversational broadcasts, and both listen and search them. Also, a web service is available for users to send their own sound files to the system for the transcription. For more specific purposes, there is a speech recognition system for radiologists under development. The first experiments with it yielded promising results (only 10% errors in real-time recognition).

During the years 1997–2002 the first Estonian text-to-speech synthesiser has been developed in the cooperation of IOC, Institute of Estonian Language (IEL) and Filosoft. It belongs to the first generation of speech synthesisers based on diphones. Every speech unit corresponds to exactly on diphone (sound-to-sound transition) in the database. The output of the synthesiser is understandable, but the sound itself is monotonous, with somewhat unnatural tone and poorly coherent. This synthesiser is adapted for use by blinds. The synthesiser is an open-source product, and may be used for non-commercial and non-military purposes [29].

IEL is developing the new corpus-based version of the synthesiser where the longer speech units have been used in addition to diphones (word and phrase).

The prize for the Best Language Deed in 2010 (established by the Ministry of Education and Research) was awarded to OÜ Jumalalaegas and the Estonian Library for the Blind for developing voice guidance tools for blinds. These tools use partially the speech synthesiser for Finnish.

Looking ahead, there will be significant changes, due to the spread of smartphones as a new platform for managing customer relationships, in addition to fixed tele-

phones, the Internet and e-mail. This will also affect how speech interaction technology is used. In the long term, there will be fewer telephone-based VUIs, and spoken language apps will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepwise improvements in the accuracy of speaker-independent speech recognition via the speech dictation services already offered as centralised services to smartphone users.

"Language recognition applications for smart phones" by Tanel Alumäe and Kaarel Kaljurand developed at Institute of Cybernetics at TUT won the grand prize of Language Deed 2011.

4.2.4 Machine Translation

The idea of using digital computers to translate natural languages can be traced back to 1946 and was followed by substantial funding for research during the 1950s and again in the 1980s. Yet **machine translation** (MT) still cannot deliver on its initial promise of providing across-the-board automated translation.

At its basic level, machine translation simply substitutes words in one natural language with words in another language.

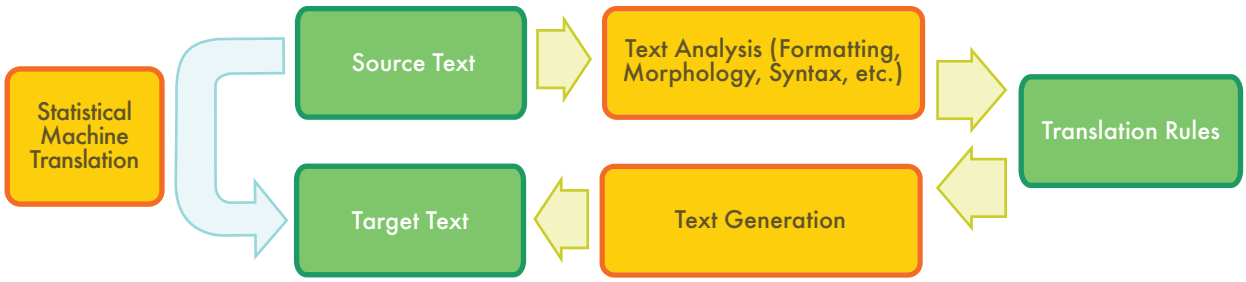
The most basic approach to machine translation is the automatic replacement of the words in a text written in one natural language with the equivalent words of another language. This can be useful in subject domains that have a very restricted, formulaic language such as weather reports. However, in order to produce a good translation of less restricted texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty is that human language is ambiguous. Ambiguity creates challenges on multiple levels, such as word sense disambiguation at the lexical level (a *mouse* is an input device of computer or an animal) or the assignment of case on the syntactic level, for example:

*Naine nägi autot ja **mees** ka.*
*Naine nägi autot ja **meest** ka.*

The woman saw the car and the man, too.

One way to build an MT system is to use linguistic rules. For translations between closely related languages, a translation using direct substitution may be feasible in cases such as the above example. However, rule-based (or linguistic knowledge-driven) systems often analyse the input text and create an intermediary symbolic representation from which the target language text can be generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by skilled linguists. This is a very long and therefore costly process. In the late 1980s when computational power increased and became cheaper, interest in statistical models for machine translation began to grow. Statistical models are derived from analysing bilingual text corpora, **parallel corpora**, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text by processing parallel versions and finding plausible patterns of words. Unlike knowledge-driven systems, however, statistical (or data-driven) MT systems often generate ungrammatical output. Data-driven MT is advantageous because less human effort is required, and it can also cover special particularities of the language (e. g., idiomatic expressions) that are often ignored in knowledge-driven systems.

The strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid ap-



6: Machine translation (left: statistical; right: rule-based)

proaches that combine both methodologies. One such approach uses both knowledge-driven and data-driven systems, together with a selection module that decides on the best output for each sentence. However, results for sentences longer than, say, 12 words, will often be far from perfect. A more effective solution is to combine the best parts of each sentence from multiple outputs; this can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

Machine Translation is particularly challenging for the Estonian language.

Machine translation is particularly challenging for the Estonian language. The potential for creating arbitrary new words by compounding makes dictionary analysis and dictionary coverage difficult; free word order and split verb constructions pose problems for analysis, also, the amount of available parallel texts is limited. In spite of this, Estonian belongs to one of the languages (currently, around 50) which can be translated by computer [30].

The history of machine translation of Estonian dates back to the end of 50s when mathematicians of University of Tartu experimented with building a system for automatic translation of mathematical texts from Russian to Estonian. The hardware used during these experiments (computer Ural) had a speed of 100 operations per second, and this mediocre performance was one of the main reasons for discontinuing the experiments.

There exist two statistical machine translation systems for Estonian. The best known is the translation service by Google. Its quality is not always satisfactory but still enables the rough understanding of the text topic and basic facts.

The research group in the University of Tartu is currently working on the idea of Estonian-English machine translation. This translation system (located on the page http://masintolge.ut.ee) allows for the translation of sentences with a limited length from Estonian to English. This system uses open source SMT decoder Moses and is trained using JRC-Acquis, OPUS and other Estonian-English parallel corpora.

The use of machine translation can significantly increase productivity provided the system is intelligently adapted to user-specific terminology and integrated into a workflow. Special systems for interactive translation support were developed, for example, at Siemens. Language portals such as the Volkswagen site provide access to dictionaries, company-specific terminology, translation memory and MT support.

There is still a huge potential for improving the quality of MT systems. The challenges involve adapting language resources to a given subject domain or user area, and integrating the technology into workflows that al-

ready have term bases and translation memories. Another problem is that most of the current systems are English-centred and only support a few languages from and into other European languages. This leads to friction in the translation workflow and forces MT users to learn different lexicon coding tools for different systems. Evaluation campaigns help to compare the quality of MT systems, the different approaches and the status of the systems for different language pairs. Figure 7 (p. 23), which was prepared during the EC Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 official EU languages (Irish was not compared). The results are ranked according to a BLEU score, which indicates higher scores for better translations [31]. A human translator would normally achieve a score of around 80 points.

The best results (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programmes and the existence of many parallel corpora (e. g., English, French, Dutch, Spanish and German). The languages with poorer results are shown in red. These languages either lack such development efforts or are structurally very different from other languages (e. g., Hungarian, Maltese, Finnish and Estonian).

4.3 OTHER APPLICATION AREAS

Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but they provide significant service functionalities "behind the scenes" of the system in question. They all form important research issues that have now evolved into individual sub-disciplines of computational linguistics. Question answering, for example, is an active area of research for which annotated corpora have been built and scientific competitions have been initiated. The concept of question answering goes beyond keyword-based searches (in which the search engine responds by delivering a collection of potentially relevant documents) and enables users to ask a concrete question to which the system provides a single answer. For example:

> *Question: How old was Neil Armstrong when he stepped on the moon?*
>
> *Answer: 38.*

While question answering is obviously related to the core area of web search, it is nowadays an umbrella term for such research issues as which different types of questions exist, and how they should be handled; how a set of documents that potentially contain the answer can be analysed and compared (do they provide conflicting answers?); and how specific information (the answer) can be reliably extracted from a document without ignoring the context.

Language technology applications often provide significant service functionalities 'behind the scenes' of larger software systems.

Question answering is in turn related to information extraction (IE), an area that was extremely popular and influential when computational linguistics took a statistical turn in the early 1990s. IE aims to identify specific pieces of information in specific classes of documents, such as the key players in company takeovers as reported in newspaper stories. Another common scenario that has been studied is reports on terrorist incidents. The task here consists of mapping appropriate parts of the text to a template that specifies the perpetrator, target, time, location and results of the incident. Domain-specific template-filling is the central characteristic of IE, which makes it another example of a "behind the scenes" technology that forms a well-demarcated research area, which in practice needs to be embedded into a suitable application environment.

Text summarisation and **text generation** are two borderline areas that can act either as standalone applications or play a supporting role. Summarisation attempts to give the essentials of a long text in a short form, and is one of the features available in Microsoft Word. It mostly uses a statistical approach to identify the "important" words in a text (i. e., words that occur very frequently in the text in question but less frequently in general language use) and determine which sentences contain the most of these "important" words. These sentences are then extracted and put together to create the summary. In this very common commercial scenario, summarisation is simply a form of sentence extraction, and the text is reduced to a subset of its sentences. An alternative approach, for which some research has been carried out, is to generate brand new sentences that do not exist in the source text.

This requires a deeper understanding of the text, which means that so far this approach is far less robust. On the whole, a text generator is rarely used as a stand-alone application but is embedded into a larger software environment, such as a clinical information system that collects, stores and processes patient data. Creating reports is just one of many applications for text summarisation.

There exist only prototype versions of these tools for Estonian. The summarisation tool for Estonian (EstSum) focuses on extraction methods from a single document. The area of texts is limited: EstSum considers that the input text consists of a formatted news text.

Question answering, information extraction, and summarisation have been the focus of numerous open competitions in the USA since the 1990s, primarily organised by the government-sponsored organisations DARPA and NIST. These competitions have significantly improved the state of the art, but their focus has mostly been on the English language. As a result, there are hardly any annotated corpora or other special resources for these tasks in Estonian. When summarisation systems use purely statistical methods, they are largely language-independent and a number of research prototypes are available. For text generation, reusable components have traditionally been limited to surface realisation modules (generation grammars) and most of the available software is for the English language.

4.4 EDUCATIONAL PROGRAMMES

Language technology is a very interdisciplinary field that involves the combined expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists among others.
Two universities in Estonia are involved in language technology education [33]:

- University of Tartu. Students of Estonian and Fenno-Ugric languages can study computer linguistics as a specialized module both during BSc and MSc program. The module includes various language theory and computer science courses, e. g., Programming for Linguists and Language Theory for Linguists. BSc and MSc students of Information Technology can study Language Technology as a specialized module. Many of these courses have been created jointly by the Department of Mathematics and Informatics and the Department of Philosophy. On the PhD level, the relevant research training is typically carried out under general linguistics or computer science.
- Tallinn University of Technology. Some PhD students of information technology are specializing in the field of speech technology, following individual study programs.

In 2009, two doctoral schools were established which involve PhD students of computer linguistics and language technology: the doctoral school of information

and communication technology (includes students of language technology), and the doctoral school of language theory, philosophy and semiotics (includes students of computer linguistics).

4.5 NATIONAL PROGRAMMES AND INITIATIVES

The language technology was actively researched in Estonia already in 1950s when the first industrial grade computers appeared in universities and research labs. However, in the early 1990s the funding of many research activities changed considerably, with some research fields falling into extinction. HLT research groups survived quite well due to successful participation in several international projects (e. g., EU Copernicus) [33]. Starting at the end of the 1990s, additional funding sources were opened:

- the Estonian Language Technology programme initiated by the Estonian Informatics Centre (1998–2000). Within this programme the first Development Plan for Estonian Language Technology was compiled in 1999;
- the national programmes "Estonian Language and Cultural Heritage" (1999–2003) and "Estonian Language and National Memory" (2004–2008) included sub-programmes for HLT.
- HLT key-players were involved also in EU FP5 project "eVikings II: Establishment of the Virtual Centre of Excellence for IST RTD in Estonia" (2002–2005).

National Programme for Estonian Language Technology (NPELT) was compiled in 2005 by a group of HLT experts and launched by the Ministry of Education and Research in 2006 for a period of five years (2006–2010). The main goal of NPELT was to develop technology support for the Estonian language to the level that would allow functioning of Estonian in the modern information society. NPELT funded HLT-related R&D activities including creation of reusable language resources and development of essential linguistic software (up to the working prototypes) as well as bringing the relevant language technology infrastructure up to date. The resources and prototypes funded by the national programme are publicly available [34].

The Center of Estonian Language Resources (CELR) spun out of a project financed by NPELT. The aim of the Centre is to create an infrastructure to make available Estonian language resources and NLP software. By the end of 2011 a consortium was formed by three partners: University of Tartu, the Institute of Cybernetics at Tallinn University of Technology and the Institute of the Estonian Language.

Continuing National Programme [35] regarding the support of Estonian language technology is going on 2011–2017. This programme will focus more on applications and on making the developed resources and tools publicly available.

The Ministry of Education and Research is funding also more research-oriented projects on language technology using targeted financing schema and grants of Estonian Science Foundation.

The Center of Excellence of Computer Science (acting from 2008 to 2015) includes computational linguists from the University of Tartu and the Institute of Cybernetics of TTU (Tallinn University of Technology).

Estonia has taken part in Common Language Resources and Technology Infrastructure (CLARIN, see http://www.clarin.eu) activities since 2008. CLARIN legal form is ERIC (European Research Infrastructure Consortium) from February 29, 2012. Center of Estonian Language Resources as an object included in the Estonian Research Infrastructures Roadmap will function as CLARIN centre of Estonia.

4.6 AVAILABILITY OF TOOLS AND RESOURCES

Figure 7 provides a rating for language technology support for the Estonian language. This rating of existing tools and resources was generated by leading experts in the field who provided estimates based on a scale from 0 (very low) to 6 (very high) using seven criteria. The key results for Estonian can be summed up as follows:

- Tools for speech recognition and speech synthesis have been developed by research institutions; however these tools can be regarded as prototypes, not as mature products.
- Although the automatic morphological analysis of Estonian is complicated, the efficiency of morphological tools (such as tokeniser, lemmatiser and morphological analyser) is comparable to similar tools for other major European languages. Since the tools have been developed as commercial products, their licensing conditions have not allowed for unrestricted use by general public. On the other hand, the quality of publicly available tools has remained fairly modest, therefore the public preference has settled for commercial utilities. Syntactic parsers with a broad coverage of Estonian have been developed only using one rule-based grammatical formalism. The grammar of the parser has been adapted for different genres of language. Nevertheless, the development of these tools must continue for achieving better performance. Semantics is more difficult to process than syntax, and text semantics is more difficult to process than word and sentence semantics. Semantic tools and resources are scored low. Thus, programs and initiatives are needed to substantially boost this area both with regard to basic research and the development of annotated corpora.
- Programs for text interpretation require extensive semantic analysis and these programs are only in the initial state of development.
- The only available tool for Estonian text generation is morphological synthesiser.
- Most of the users prefer Google machine translation service. Also, Estonian-English machine translation system is under development in the University of Tartu. There would probably be a high demand for the Russian-Estonian and Estonian-Russian automatic translation service.
- With regard to resources such as reference corpora, lexicons, wordnets and terminologies, the situation is also reasonably good for Estonian since substantial resources have been built in recent decades. While some reference corpora of high quality exist, syntactically and semantically annotated corpora are small in size. The work on multimodal corpora has begun recently.
- Research was successful in designing particular high quality software, but many of the resources lack standardisation and detail documentation, i. e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.

To sum up, the results indicate that Estonian stands reasonably well with respect to the most basic language technology tools and resources, such as tokenisers, PoS taggers, morphological analysers, shallow parsers, reference corpora, treebanks and tools for speech technology. Furthermore, there exist some tools for summarisation, machine translation, as well as resources like parallel corpora, and specialized corpora. However, these tools and resources are rather simple and have a limited functionality for some of the areas. For instance, parallel corpora only exist for very few language pairs and for limited text genres. When it comes to more advanced

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	2	5	2.8	2.8	3	3	3
Speech Synthesis	2	5	2.8	2.8	3	2	3
Grammatical analysis	2.5	3.5	3.2	2.8	4	2.5	3.5
Semantic analysis	1	1.3	0.9	0.9	1.3	1.3	1.7
Text generation	0	0	0	0	0	0	0
Machine translation	3	3	1.4	2.1	3	4	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	3	5	2.5	2.1	3	2.5	
Speech corpora	2	5	2.1	2.8	4	4	4
Parallel corpora	2	2	2.1	1.4	3	3	2
Lexical resources	3.5	4	3.2	2.8	3.5	3.5	3.5
Grammars	2	5	2.8	2.8	3	3	3

7: State of language technology support for Estonian

fields like text semantics, language generation, and annotated multimodal data, Estonian clearly lacks basic tools and resources even if some of these are currently under development. The research regarding the most advanced tools and resources like discourse processing, dialogue management, semantics and discourse corpora has yielded first results but the resources need still to complement in amount and quality and tools have a quite limited scope. Most of these tools (except morphological analyser) have been developed by research institutions and can be regarded as prototypes, not as mature products. The development was supported by different national programmes of language technology what guarantees the availability of these tools.

4.7 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section presents an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using a five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

Figures 8 to 11 show that, although the government has increased the support to the Estonian LT recent years, Estonian language belongs to the fourth or fifth clusters. The results of Estonian and Finnish are quite similar and Estonian is slightly better equipped than most other languages with a similar number of speakers, such as Latvian, Lithuanian, Icelandic and Maltese. All these languages lag far behind large languages like German and French, for instance. But even LT resources and tools for those languages clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications

For building more sophisticated applications, such as machine translation, there is a clear need for resources and technologies that cover a wider range of linguistic aspects and allow a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced application areas, including high-quality machine translation.

4.8 CONCLUSIONS

In this series of white papers, we have made an important effort by assessing the language technology support for 30 European languages, and by providing a high-level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and its related stakeholders are now in a position to design a large scale research and development programme aimed at building a truly multilingual, technology-enabled communication across Europe.

The results of this white paper series show that there is a dramatic difference in language technology support between the various European languages. While there are good quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources. Others have basic tools and resources but the implementation of for example semantic methods is still far away. Therefore a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation.

In the case of the Estonian language, we can be cautiously optimistic about the current state of language technology support. Supported by larger research programs in the past, there exists a language technology research scene in Estonia. Unfortunately, the industry consists only of a few SMEs.

For Estonian, a number of technologies and resources exist, but far less than for English. Language technology support today is by far not in a state that is needed for offering the support a true multilingual knowledge society needs.

The need for large amounts of data and the extreme complexity of language technology systems makes it vital to develop a new infrastructure and a more coherent research organisation to spur greater sharing and cooperation.

Finally there is a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. In addition, there is an overall lack of coordination with programmes in other EU countries and at the European Commission level.

The long term goal of META-NET is to enable the creation of high-quality language technology for all languages. This requires all stakeholders - in politics, research, business, and society - to unite their efforts. The resulting technology will help tear down existing barriers and build bridges between Europe's languages, paving the way for political and economic unity through cultural diversity.

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

8: Speech processing: state of language technology support for 30 European languages

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

9: Machine translation: state of language technology support for 30 European languages

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

10: Text analysis: state of language technology support for 30 European languages

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 Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese

11: Speech and text resources: State of support for 30 European languages

5

ABOUT META-NET

META-NET is a Network of Excellence funded by the European Commission [36]. The network currently consists of 54 members from 33 European countries. META-NET forges **META**, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots. Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language re-sources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

KIRJANDUS REFERENCES

[1] Aljoscha Burchardt, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] Riigiportaal. https://www.eesti.ee/est/teemad/ettevotja/riigiportaali_abi/partnerile_1/lemmatiseerja.

[3] Aljoscha Burchardt, Georg Rehm, and Felix Sasaki. The Future European Multilingual Information Society – Vision Paper for a Strategic Research Agenda, 2011. http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf.

[4] Directorate-General Information Society & Media of the European Commission. User Language Preferences Online, 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[5] European Commission. Multilingualism: an Asset for Europe and a Shared Commitment, 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[6] Directorate-General of the UNESCO. Intersectoral Mid-term Strategy on Languages and Multilingualism, 2007. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

[7] Directorate-General for Translation of the European Commission. Size of the Language Industry in the EU, 2009. http://ec.europa.eu/dgs/translation/publications/studies.

[8] Estonian Institute. Estonian in a world context. http://www.estonica.org/en/Society/The_Estonian_Language/Estonian_in_a_world_context/.

[9] Anastassia Zabrodskaja. Eesti keel Eestis teise keelena. http://www.emakeeleselts.ee/vaatmikud/eesti_keel_teise_keelena.pdf, 2008. In Estonian. English title: Estonian as a Second Language in Estonia.

[10] Eesti keelenõukogu. Eesti keele arendamise strateegia 2004-2010. http://eki.ee/keelenoukogu/strat_et.pdf. In Estonian. English title: Development Strategy of the Estonian Language 2004–2010.

[11] Estonian Institute. Estonian Sign Language. http://www.estinst.ee/publications/language/sign.html.

[12] Helle Metslang. Estonian grammar between Finnic and SAE: some comparisons. In *Language Typology and Universals*, pages 49–71, 2009.

[13] Mati Erelt, Tiiu Erelt, and Kristiina Ross. *Eesti keele käsiraamat*. Eesti Keele Sihtasutus, Tallinn, 2007. In Estonian. English title: The handbook of Estonian language.

[14] Mati Erelt. Estonian language. In *Linguistica Uralica Supplementary Series*, volume 1, 2003.

[15] Estonian Language Council. Development plan for Estonian language (2011-2017). http://ekn.hm.ee/system/files/Eesti_keele_arengukava_2011-2017_0.pdf.

[16] Estonian Institute. Eesti keel ja kultuur maailmas — Ülikoolid (universities). http://ekkm.estinst.ee/keskused/ylikoolid/.

[17] Eesti Statistikaamet. Kolmveerand Eesti elanikest kasutab internetti. http://www.stat.ee/dokumendid/47013, 2010. In Estonian. English title: Three-fourths of Estonian population uses the internet.

[18] neti.ee. http://www.neti.ee/cgi-bin/teema/INFO_JA_MEEDIA/Ajalehed/.

[19] Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer, and Ralf Klabunde, editors. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Spektrum Akademischer Verlag, 2009. In German. English title: *Computational Linguistics and Language Technology: An Introduction*.

[20] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice Hall, 2009.

[21] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[22] Language Technology World (LT World). http://www.lt-world.org/.

[23] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing)*. Cambridge University Press, 1998.

[24] Jerrold H. Zar. Candidate for a Pullet Surprise. *Journal of Irreproducible Results*, page 13, 1994.

[25] Filosoft OÜ. http://www.filosoft.ee/index_en.html.

[26] Spiegel Online. Google zieht weiter davon (Google is still leaving everybody behind), 2009. http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.

[27] Juan Carlos Perez. Google Rolls out Semantic Search Capabilities, 2009. http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html.

[28] Laboratory of Phonetics and Speech Technology Institute of Cybernetics at TTU. http://www.phon.ioc.ee/dokuwiki/doku.php?id=software:software.en.

[29] Eesti Keele Instituut. Eesti keele kõnesüntees. http://www.eki.ee/keeletehnoloogia/projektid/syntees/. In Estonian. English title: Speech Synthesis for Estonian Language.

[30] Heiki-Jaan Kaalep and Mare Koit. Kuidas masin tõlgib. *Keel ja Kirjandus*, pages 726–738, 10 2010. In Estonian. English title: How Does a Machine Translate.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA, 2002.

[32] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 Machine Translation Systems for Europe. In *Proceedings of MT Summit XII*, 2009.

[33] Einar Meister, Tiit Roosmaa, and Jaak Vilo. Estonian language technology anno 2009. In R. Domeij, K. Koskenniemi, S. Krauwer, B. Maegaard, E. Rögnvaldsson, and K de Smedt, editors, *Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Common Language Resources*, number 5 in NEALT Proceedings Series, pages 21–26. University of Tartu, 2009.

[34] National Programme for Estonian Language Technology. Riiklik programm "Eesti keele keeletehnoloogiline tugi (2006–2010)". http://www.keeletehnoloogia.ee/ekktt-1.

[35] National Programme for Estonian Language Technology. Riiklik programm "Eesti keeletehnoloogia (2011–2017)". http://www.keeletehnoloogia.ee.

[36] Georg Rehm and Hans Uszkoreit. Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):51–52, April/May 2011.

[37] Estonian Language Council. Development strategy of the Estonian language 2004–2010. http://eki.ee/keelenoukogu/strat_en.pdf.

B

META-NETI LIIKMED META-NET MEMBERS

Austria	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Belgia	Belgium	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp: Walter Daelemans
		Centre for Processing Speech and Images, University of Leuven: Dirk van Compernolle
Bulgaaria	Bulgaria	Institute for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
Eesti	Estonia	Institute of Computer Science, University of Tartu: Tiit Roosmaa, Kadri Vider
Hispaania	Spain	Barcelona Media: Toni Badia, Maite Melero
		Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel
		Aholab Signal Processing Laboratory, University of the Basque Country: Inma Hernaez Rioja
		Centre for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
		Department of Signal Processing and Communications, University of Vigo: Carmen García Mateo
Horvaatia	Croatia	Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb: Marko Tadić
Iirimaa	Ireland	School of Computing, Dublin City University: Josef van Genabith
Island	Iceland	School of Humanities, University of Iceland: Eiríkur Rögnvaldsson
Itaalia	Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Language Technology Research Unit, Fondazione Bruno Kessler: Bernardo Magnini
Kreeka	Greece	R.C. "Athena", Institute for Language and Speech Processing: Stelios Piperidis
Küpros	Cyprus	Language Centre, School of Humanities: Jack Burston
Leedu	Lithuania	Institute of the Lithuanian Language: Jolanta Zabarskaitė
Läti	Latvia	Tilde: Andrejs Vasiļjevs
		Institute of Mathematics and Computer Science, University of Latvia: Inguna Skadiņa
Luksemburg	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Madalmaad	Netherlands	Utrecht Institute of Linguistics, Utrecht University: Jan Odijk

		Computational Linguistics, University of Groningen: Gertjan van Noord
Malta	Malta	Department Intelligent Computer Systems, University of Malta: Mike Rosner
Norra	Norway	Department of Linguistic, Literary and Aesthetic Studies, University of Bergen: Koenraad De Smedt
		Department of Informatics, Language Technology Group, University of Oslo: Stephan Oepen
Poola	Poland	Institute of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
		University of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Department of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University: Zygmunt Vetulani
Portugal	Portugal	University of Lisbon: António Branco, Amália Mendes
		Spoken Language Systems Laboratory, Institute for Systems Engineering and Computers: Isabel Trancoso
Prantsusmaa	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur and Institute for Multilingual and Multimedia Information: Joseph Mariani
		Evaluations and Language Resources Distribution Agency: Khalid Choukri
Rootsi	Sweden	Department of Swedish, University of Gothenburg: Lars Borin
Rumeenia	Romania	Research Institute for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiș
		Faculty of Computer Science, University Alexandru Ioan Cuza of Iași: Dan Cristea
Saksamaa	Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
		Human Language Technology and Pattern Recognition, RWTH Aachen University: Hermann Ney
		Department of Computational Linguistics, Saarland University: Manfred Pinkal
Serbia	Serbia	University of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
		Pupin Institute: Sanja Vranes
Slovakkia	Slovakia	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Sloveenia	Slovenia	Jožef Stefan Institute: Marko Grobelnik
Soome	Finland	Computational Cognitive Systems Research Group, Aalto University: Timo Honkela
		Department of Modern Languages, University of Helsinki: Kimmo Koskenniemi, Krister Lindén
Suurbritannia	UK	School of Computer Science, University of Manchester: Sophia Ananiadou

		Institute for Language, Cognition and Computation, Centre for Speech Technology Research, University of Edinburgh: Steve Renals
		Research Institute of Informatics and Language Processing, University of Wolverhampton: Ruslan Mitkov
Šveits	Switzerland	Idiap Research Institute: Hervé Bourlard
Taani	Denmark	Centre for Language Technology, University of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Tšehhi Vabariik	Czech Republic	Institute of Formal and Applied Linguistics, Charles University in Prague: Jan Hajič
Ungari	Hungary	Research Institute for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics: Géza Németh, Gábor Olaszy

Ligi 100 keeletehnoloogiaeksperti – META-NETis osalevate maade ja keelte esindajad – arutasid läbi ja viimistlesid Valge raamatu sarja põhitulemusi META-NETi koosolekul Berliinis 21. ja 22. oktoobril 2011.– About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

META-NETI VALGE RAAMATU SARI

THE META-NET WHITE PAPER SERIES

baski	Basque	euskara
bulgaaria	Bulgarian	български
eesti	Estonian	eesti
galiitsia	Galician	galego
hispaania	Spanish	español
hollandi	Dutch	Nederlands
horvaadi	Croatian	hrvatski
iiri	Irish	Gaeilge
inglise	English	English
islandi	Icelandic	íslenska
itaalia	Italian	italiano
katalaani	Catalan	català
kreeka	Greek	ελληνικά
leedu	Lithuanian	lietuvių kalba
läti	Latvian	latviešu valoda
malta	Maltese	Malti
norra Bokmål	Norwegian Bokmål	bokmål
norra Nynorsk	Norwegian Nynorsk	nynorsk
poola	Polish	polski
portugali	Portuguese	português
prantsuse	French	français
rootsi	Swedish	svenska
rumeenia	Romanian	română
saksa	German	Deutsch
serbia	Serbian	српски
slovaki	Slovak	slovenčina
sloveeni	Slovene	slovenščina
soome	Finnish	suomi
taani	Danish	dansk
tšehhi	Czech	čeština
ungari	Hungarian	magyar

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the Estonian language. It is part of a series that analyses the available language resources and technologies for 30 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations, non-governmental organisations, language communities and European universities. META-NET's vision is high-quality language technology for all European languages.

Oma igapäevases suhtluses puutuvad Euroopa kodanikud, äripartnerid ja poliitikud paratamatult kokku keelebarjääridega. Keeletehnoloogia suudab neid barjääre ületada ning võimaldada uudseid liideseid tehnoloogiatele ja teadmistele. Käesolev valge raamat esitab eesti keele keeletehnoloogilise toe hetkeseisu, moodustades osa seeriast, mis analüüsib 30 Euroopa keele olemasolevaid keeleressursse ja -tehnoloogiaid. Selle analüüsi viis läbi Euroopa Komisjoni rahastatud tippteadmiste võrgustik META-NET. META-NET koosneb 33 riigi 54 uurimiskeskusest, mis teevad koostööd ärimaailma, valitsus- ja teadusasutuste, valitsusväliste organisatsioonide, keelekogukondade ja Euroopa ülikoolide huvitatud osapooltega. META-NETi eesmärk on saavutada kõrgekvaliteediline keeletehnoloogia kõigile Euroopa keeltele.

"If we do not implement the development plan for language technology or do not cooperate with other countries in the same direction, in future Estonian will [...] be marginalized in information society."
– Development Plan of the Estonian Language 2011–2017

"Kas eesti keelt on võimalik kasutada kõige moodsamas tehnoloogias ja kas arvuti on võimeline suhtlema kujundlikus eesti keeles – see määrab meie keele säilimise tuleviku maailmas."
– Tõnis Lukas, haridus- ja teadusminister 2007–2011, EV90 raames korraldatud Keeletalgutel