META NET

White Paper Series Valkoiset kirjat

# THE FINNISH LANGUAGE IN THE DIGITAL AGE

# SUOMEN KIELI DIGITAALISELLA AIKAKAUDELLA

Kimmo Koskenniemi
Krister Lindén
Lauri Carlson
Martti Vainio
Antti Arppe
Mietta Lennes
Hanna Westerlund
Mirka Hyvärinen
Imre Bartis
Pirkko Nuolijärvi
Aino Piehl

Springer

White Paper Series Valkoiset kirjat

# THE FINNISH LANGUAGE IN THE DIGITAL AGE

# SUOMEN KIELI DIGITAALISELLA AIKAKAUDELLA


Kimmo Koskenniemi Helsingin yliopisto
Krister Lindén Helsingin yliopisto
Lauri Carlson Helsingin yliopisto
Martti Vainio Helsingin yliopisto
Antti Arppe Helsingin yliopisto
Mietta Lennes Helsingin yliopisto
Hanna Westerlund Helsingin yliopisto
Mirka Hyvärinen Helsingin yliopisto
Imre Bartis Helsingin yliopisto
Pirkko Nuolijärvi KOTUS
Aino Piehl KOTUS

Georg Rehm, Hans Uszkoreit
(toimittajat, editors)

*Editors*
Georg Rehm
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: georg.rehm@dfki.de

Hans Uszkoreit
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: hans.uszkoreit@dfki.de

ISSN 2194-1416 ISSN 2194-1424 (electronic)
ISBN 978-3-642-27247-9 ISBN 978-3-642-27248-6 (eBook)
DOI 10.1007/978-3-642-27248-6
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940571



Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# ESIPUHE PREFACE

META-NET Valkoiset kirjat -julkaisusarjan tavoitteena on edistää tietämystä kieliteknologiasta ja sen tarjoamista mahdollisuuksista. Tämä julkaisu haluaa herättää opettajia, toimittajia, poliitikkoja, kieliyhteisöjä ja muitakin.

Euroopan kielten kieliteknologisten sovellusten saatavuus vaihtelee. Niinpä myös toimenpiteet, joita jatkossa tarvitaan tukemaan kieliteknologioiden tutkimusta ja kehitystä, ovat eri kielten kohdalla erilaisia ja riippuvat kielen ominaispiirteistä ja kieliyhteisön koosta.

Euroopan komission rahoittaman META-NET -huippuosaamisverkoston kartoitustyö tässä valkoisten kirjojen sarjassa (p. 81) kattaa Euroopan 23 virallisen kielen sekä tärkeiden kansallisten ja paikallisten kielten kieliaineistot ja kieliteknologiat. Tulosten perusteella kaikkien kartoitettujen kielten tutkimus kärsii merkittävästä resurssien puutteesta. Yksityiskohtaisempi nykyisen tilanteen selvitys vahvistaa tulevan tutkimuksen vaikutusta ja vähentää riskejä.

META-NET koostuu 33 valtion 54 tutkimuskeskuksesta [1] (s. 77), jotka tekevät yhteistyötä useiden toimijoiden ja intressiryhmien kanssa. Mukana on liikeyrityksiä, julkisen hallinnon yksiköitä, teollisuuden edustajia, tutkimusyksiköitä, tietotekniikan alan yrityksiä, teknologian tuottajia ja eurooppalaisia yliopistoja. Työn tuloksena on syntymässä teknologinen visio osana strategista tutkimuslinjausta osoittamaan, miten kieliteknologiat auttavat Euroopan tutkimusyhteisöä ratkaisemaan keskeisiä tutkimuskysymyksiä vuoteen 2020 mennessä.

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. 81). The analysis focuses 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 future research.

META-NET consists of 54 research centres in 33 European countries [1] (p. 77). META-NET is working with stakeholders from economy (software companies, technology providers and 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.

Tämän raportin tekijät ovat kiitollisia saksankielisen META-NET valkoisen kirjan tekijöille luvasta käyttää raporttinsa kielestä riippumattomien osioiden tekstejä osana tämän raportin englanninkielistä osuutta sekä lähteenä suomenkieliselle käännökselle [2].

Tämän valkoisen kirjan tuottamiseen on myönnetty rahoitusta Euroopan komission seitsemännestä puiteohjelmasta ja tieto- ja viestintäteknologioiden tukiohjelmasta seuraavien sopimusten perusteella T4ME (rahoitussopimus 249119), CESAR (rahoitussopimus 271022), METANET4U (rahoitussopimus 270893) ja META-NORD (rahoitussopimus 270899).

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 [2].

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 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

# SISÄLLYSLUETTELO TABLE OF CONTENTS

## SUOMEN KIELI DIGITAALISELLA AIKAKAUDELLA

# THE FINNISH LANGUAGE IN THE DIGITAL AGE

# TIIVISTELMÄ

Tietotekniikka muuttaa jokapäiväistä elämäämme. Käytämme tietokoneita kirjoittamiseen, tekstin muokkaamiseen, laskemiseen, tiedon etsimiseen ja yhä enemmän myös lukemiseen, musiikin kuunteluun sekä valokuvien ja elokuvien katseluun. Kannamme taskuissamme pieniä tietokoneita, joilla soitamme puheluja, lähetämme sähköpostia ja viihdytämme itseämme siellä missä kulloinkin satumme olemaan. Kuinka tämä valtava informaation, tietämyksen ja arkisen viestinnän digitalisoituminen vaikuttaa kieleemme? Muuttuuko suomen kieli tai voiko se jopa kadota? Kaikki tietokoneemme ovat yhteydessä toisiinsa entistä tiheämmän ja tehokkaamman maailmanlaajuisen verkon kautta. Tyttö Ipanemassa, tullimies Imatralla ja insinööri Katmandussa voivat jutella ystäviensä kanssa Facebookissa, mutta toisiinsa he tuskin koskaan verkossa törmäävät. Jos he ovat huolissaan korvasärystä, he käyvät lukemassa Wikipediasta kaiken mahdollisen tämän vaivan hoitoon liittyvän, mutteivät silloinkaan lue samaa artikkelia. Ja kun Euroopan nettikansalaiset keskustelevat Fukushiman ydinonnettomuuden vaikutuksista eurooppalaiseen energiapolitiikkaan, tapahtuu ajatustenvaihto erikseen kunkin kieliyhteisön sisäisillä keskustelupalstoilla. Kielet erottavat edelleenkin sen minkä Internet voisi yhdistää. Tyydymmekö tähän tilanteeseen myös tulevaisuudessa?

Tieteiselokuvissa kaikki puhuvat samaa kieltä. Voisiko tämä yhteinen kieli olla suomi, vaikka astronautit harvoin lausuvat suomalaisia sanoja yhtä luonnollisesti kuin he puhuvat englantia? Monet maailman 6000 kielestä eivät tule selviytymään globalisoituneessa digitaalisessa tietoyhteiskunnassa. Arviolta vähintään 2000 kieltä on tuomittu sukupuuttoon tulevina vuosikymmeninä. Joitakin kieliä mahdollisesti käytetään jatkossakin perheissä ja kyläyhteisöissä, mutta ei yrityksissä tai akateemisessa maailmassa. Minkälaiset siis ovat suomen kielen selviytymismahdollisuudet?

Suomea puhuu yli 5 miljoonaa ihmistä, joten se on moniin muihin kieliin verrattuna kohtalaisen hyvässä asemassa. Suomenkielisiä julkisia televisiokanavia on neljä ja yksityisiä yli 30. Useimmat kansainväliset elokuvat tekstitetään suomeksi. Suomen kieli on todennäköisesti hieman vahvistanut asemiaan sen jälkeen kun Suomi liittyi EU:n täysjäseneksi. Kielen puhujien, kirjojen, elokuvien ja televisiokanavien määrän lisäksi tietyn kielen tilanne riippuu myös sen digitaalisesta läsnäolosta tietoverkoissa ja sovellusohjelmissa. Tälläkin mittapuulla suomi sijoittuu kohtalaisen hyvin: kaikki keskeiset kansainväliset ohjelmistotuotteet ovat saatavilla suomalaisina versioina, suomenkielisessä Wikipediassa on yli 290 000 artikkelia ja verkkotunnus .fi on hyvin suosittu.

Kieliteknologian alalla suomen kielelle on tarjolla kohtuullinen määrä tuotteita, teknologioita ja kielivaroja. On olemassa suomenkielisiä sovelluksia ja työkaluja puhesynteesiä, puheentunnistusta, tiedonhakua sekä oikeinkirjoituksen ja kieliopin tarkistusta varten. On olemassa myös joitakin automaattista kääntämistä varten kehitettyjä sovelluksia, vaikka ne eivät usein tuotakaan kielellisesti ja idiomaattisesti oikeita käännöksiä varsinkaan kun suomi on kohdekielenä. Tähän ovat osittain syynä suomen kielen erityispiirteet.

Tieto- ja viestintätekniikka valmistautuvat nyt seuraavaan vallankumoukseen. Mikrotietokoneita, multimediaa, tietoverkkoja, laitteiden pienentymistä, multimediaa, mobiililaitteita ja pilvilaskentaa seuraava teknologian sukupolvi luo ohjelmistoja, jotka ymmärtävät kirjainten ja äänteiden lisäksi myös kokonaisia sanoja ja lauseita. Tällaiset ohjelmistot palvelevat käyttäjiään entistä paremmin, koska ne puhuvat ja ymmärtävät heidän kieltään. Alan edelläkävijöitä ovat ilmainen onlinepalvelu Google Translate, joka kääntää 57 kielen välillä, IBM:n supertietokone Watson, joka päihitti Jeopardy-tietovisassa Yhdysvaltojen mestarin, sekä Applen iPhoneen kehittämä Siri-avustaja, joka reagoi äänikomentoihin ja vastaa englanniksi, saksaksi, ranskaksi ja japaniksi esitettyihin kysymyksiin.

Tietotekniikan seuraava sukupolvi tulee hallitsemaan ihmiskielen niin laajasti, että erikieliset käyttäjät pystyvät viestimään keskenään kukin omalla kielellään. Helppokäyttöisten äänikomentojen pohjalta laitteet osaavat hakea automaattisesti tärkeimmät uutiset ja muuta tietoa maailman digitaalisista tietovarannoista. Kieliteknologian avulla voidaan tehdä automaattisia käännöksiä ja avustaa tulkkeja. Sitä voi käyttää tulevaisuudessa myös keskustelujen ja asiakirjojen tiivistämiseen sekä opiskelun tukena. Kieliteknologia voi esimerkiksi auttaa maahanmuuttajia oppimaan suomea ja integroitumaan paremmin suomalaiseen kulttuuriin.

Seuraavan sukupolven tieto- ja viestintätekniikan avulla kehitellään jo nyt tutkimuslaboratorioissa teollisuuden ja palvelualan robotteja, jotka sekä ymmärtävät täysin mitä käyttäjät niiltä haluavat että osaavat raportoida omista saavutuksistaan. Tällaiseen suoritustasoon pääseminen vaatii paljon enemmän kuin pelkkien merkistöjen, sanakirjojen, oikolukuohjelmien ja ääntämissääntöjen käyttöä. Yksinkertaistettu lähestymistapa teknologiassa ei enää riitä, vaan on ryhdyttävä mallintamaan kieltä kokonaisvaltaisesti. On samanaikaisesti huomioitava sekä syntaksi että semantiikka, jotta myös mutkikkaita kysymyksiä voidaan ymmärtää ja antaa niihin perusteellisia ja relevantteja vastauksia.

Englannin ja suomen välillä on kuitenkin ammottava teknologinen kuilu, joka tätä nykyä vieläpä levenee. 1980- ja 1990-luvun menestyksekkäiden tutkimussaavutusten jälkeen Suomi on nyt menettämässä rooliaan kieliteknologian edistäjänä. Kieliteknologian perustutkimusta rahoitettiin tutkimuksen huippuyksikön tasolla 1980- ja 1990-luvuilla, mikä johti useiden kehitettyihin tuotteisiin perustuvien yritysten perustamiseen.

Perustutkimuksen rahoituksen kauden jälkeen teknologiateollisuuteen liittyvät hankkeet ovat saaneet vain pienimuotoista rahoitusta Tekesiltä (teknologian ja innovaatioiden kehittämiskeskukselta). Tämän seurauksena Suomi (ja koko Eurooppa) menetti joitakin erittäin lupaavia huipputekniikan innovaatioita Yhdysvaltoihin, jossa tutkimuksen strateginen suunnittelu on pitkäjänteisempää ja rahoitusta on paremmin saatavilla myös uusien teknologioiden markkinoille tuomiseen. Vaikka uraauurtavalla tuoteidealla onnistuisikin saamaan varaslähdön teknologisten innovaatioiden kilpailussa, voi oman etulyöntiasemansa varmistaa vain siinä tapauksessa, että pystyy myös ylittämään maaliviivan. Muuten käteen jää pelkkä kunniamaininta Wikipediassa.

Kun kieliteknologian perustutkimuksen rahoitus väheni, siirtyivät monet suomalaiset asiantuntijat erilaisiin pienyrityksiin. Yhdysvaltalaiset yritykset käyttivät resurssejaan kehittääkseen teknologioista itselleen käyttökelpoisia tuotteita. Tästä huolimatta Suomessa on edelleen hyvin suuri tutkimuspotentiaali. Kansainvälisesti tunnettujen tutkimuskeskusten ja yliopistojen lisäksi täällä on myös innovatiivisia pieniä ja keskikokoisia kieliteknologiayrityksiä, jotka pysyvät hengissä silkan luovuuden ja valtavien ponnistusten ansiosta, vaikka niillä ei olekaan riskipääomaa tai jatkuvaa julkista rahoitusta. Suomenkielisen kieliteknologian varhaisen kaupallisen menestyksen takia ei tutkimusyhteisö enää päässyt kään käyttämään suomen kielen käsittelyyn kehitettyjä