

Language Technology, Research and Society

*”Nyttan av språkteknologi i samhället med betoning
på forskning och utveckling”*

Krister Lindén,
University of Helsinki
FIN-CLARIN

Topics

- Language Technology
- Trends and Visions on the EU level
- Distributing Language Resources

Background

- Ideas collected and proposed in META-NET
 - European Commission Network of Excellence
 - 60 research institutes from 34 countries
- Goals Reflected in
 - Strategic Research Agenda (2014-2020)
 - Horizon 2020

LANGAUAGE TECHNOLOGY

Where do end-users meet LT?

- Spelling and grammar correction
- Mono and multi-lingual search
- Recommendations for similar items
- Translation services on the web
- Speech-guided and speaking applications in mobile devices

Market size

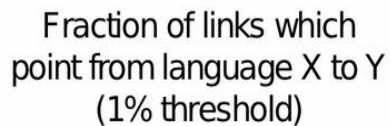
- The language industry in EU (translation, interpretation, localization and globalization) was estimated to 8.4 billion euro in 2008 and the expected growth was then 10% annually
- Even with a 5% annual growth, the market size should be approx. 11.7 billion euro in 2015
- Already today the daily throughput of Google Translate is equal to what all human translators translate annually

TRENDS AND VISIONS

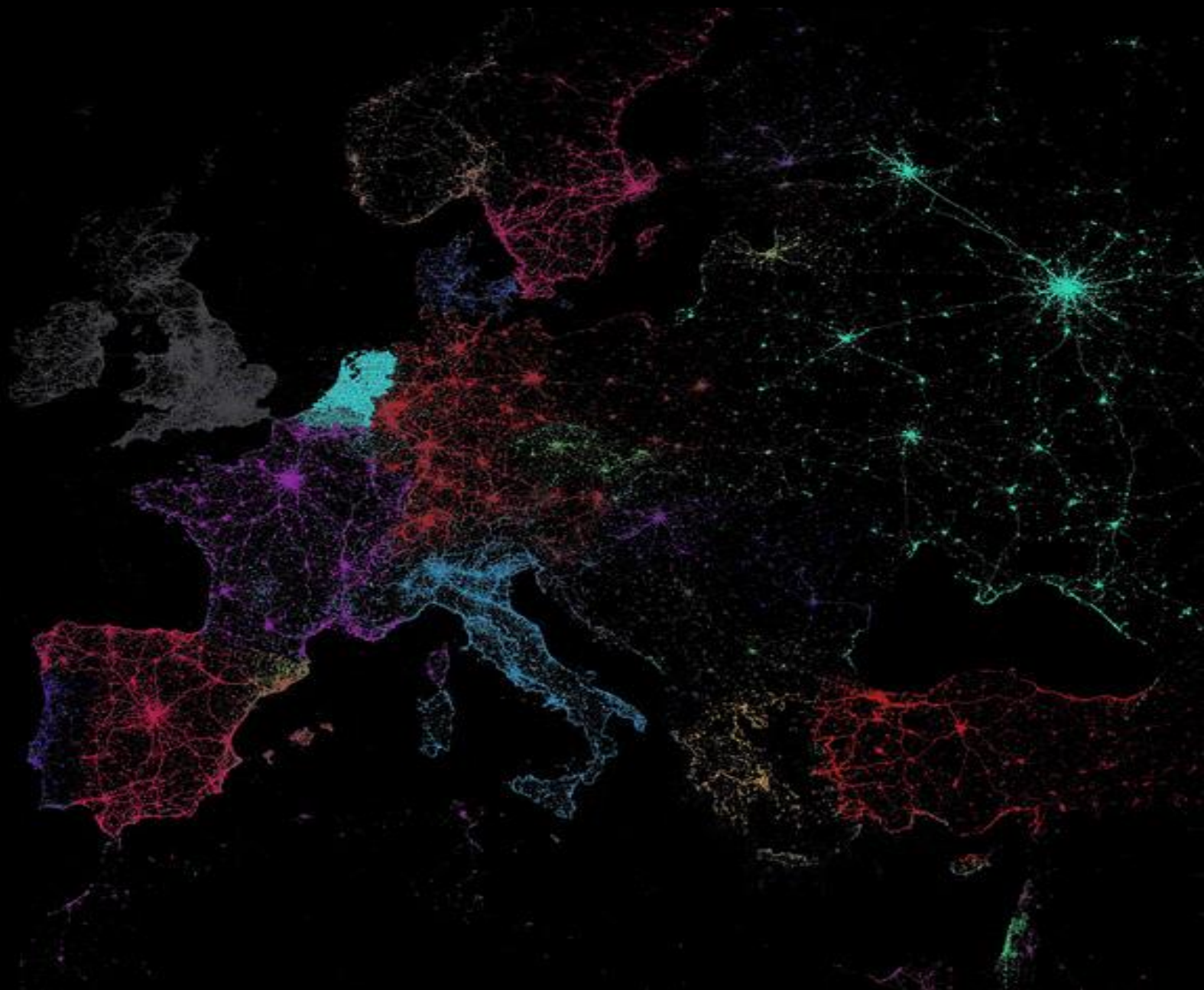
Why do we need LT?

- The most significant trade barrier is that people do not understand one another
- In EU the greatest obstacle for Information and Communication Technology (ICT) uptake is language differences, as every language area also tend to be a separate market area
- Solutions:
 - Every person needs to speak a common language (English?) or
 - We use language technology to remove the barriers

Daniel Ford, Josh Batson



- English
- Portuguese
- Indonesian
- Spanish
- Malay
- Japanese
- Dutch
- Korean
- Filipino
- Russian
- French
- Thai
- Italian
- German
- Turkish
- Arabic
- Swedish
- Danish
- Finnish
- Catalan
- Chinese
- Romanian
- Norwegian
- Lithuanian
- Slovak
- Czech
- Vietnamese
- Greek
- Hungarian
- Polish
- Afrikaans
- Slovenian
- Albanian
- Latvian
- Chinese (TW)
- Galician
- Swahili
- Hebrew
- Croatian
- Bulgarian



Horizon 2020 Visions

- **Translation Cloud** – we need services for producing immediate and reliable speech and text translation
- **Social intelligence and e-participation** – we need technology to increase understanding and dialog between and inside communities
- **Socially-aware interaction assistants** – we need technology that is capable of interpreting and reacting to other than spoken communication

Translation Drivers

- 59% of the web shops are capable of providing services in more than one language
- 81% of the internet users think that web pages in their own country should be multi-lingual
- 44% of the users in Europe think they miss some interesting information because they find web pages in a language they do not understand

Population trends

- Aging population (20-25 %): "Where did I leave my glasses?"
- Disabled persons (10 %) need equal opportunities to participate, e.g. we need sign language translation
- Migrant persons (5-15%) need language training, e.g. Swedish doctors and nurses in Norway

General Trends

- Web users need summarization and answering services, cf. Watson and Jeopardy
- Virtual interaction and augmented reality, cf. web games, netmeetings, Google glass, etc.
- Social media enable global communication and local disintegration, cf. spread of extreme ideologies

RESEARCH AND DEVELOPMENT

Data and metadata

- When the amount of data increases, we need to develop methods for linking, classifying and annotating information automatically
- Annotated corpora with pictures, sound and video (lat.csc.fi)

FirefoxIMDI BrowserTrova Search ApplicationMPI - ANNEX Interface

lat.csc.fi/ds/annex/runLoader?sessionId=6379B33C0AC6E7F7376E286A93417F76&nodeid=MPI10%23&time=280&duration=970&tiername=K-SpchGoogle

Annex 1.2.35420

manual?embed

Show tooltips

Compact

Spacious

user: anonymouslogin

Text

Grid


Subtitle

Waveform

Timeline

Combined

Video display



⏮⏪⏩⏭⏮⏪⏩⏭

FullBuffer

Information

GeneralSessionTechnical

Resource: elan-example1.eaf
Media file: elan-example1.mp4
Elapsed time: 00:00:06:360

Selected chunk:
Begin time: 00:00:05:720
End time: 00:00:06:360
Text: stroke

Mini Data Frame

Tier: none

Font size: 14

Play selection

Clear selection

Create bookmark

<|>|<><><>+ -

Play screen by screen

Play continually

Tier text font: Arial Unicode MS

Timeline

00:00:04:50000:00:05:00000:00:05:50000:00:06:00000:00:06:50000:00:07:00000:00:07:50000:00:08:000

K-Spch																	
W-Spch	and then you go the other, Saint Anna Straat to this to the center of the town, to this big rotunde.																
W-Words	and	then	you	go	the	other	Saint	Anna	Straat	to	this	to	the	center			
W-POS	con	adv	pro	v	art	adj	n	n	n	prep	dem	prep	art	n			
W-IPA	ənd ðen ju: ɡəʊ ði ɒðə sɑnt ʌnə strɑ:t tu ðɪs to ðə sɑntə əf ðə taʊn tu ðɪs bɪɡ rɒtʊndə																
W-RGU																	
W-RGph	preparation					stroke					hold		part.retraction			holc preparation stroke	
W-RGMe						Going along St. Anna Street										Going al	
K-RGU																	
K-RGph																	
K-RGMe																	

17

Firefox IMDI Browser Trova Search Application MPI - ANNEX Interface

lat.csc.fi/ds/trova/search.jsp?jsessionid=7D78548F7889D8E7B41AA6B71E0D1752&nodeid=MPI2%23&row=4

TROVA 1.4.35374 help user: anonymous login [logout]

Substring Search Single Layer Search Multiple Layer Search

Types: ☒ EAF (1)

Domain: PeWi corpus

Find here

Ready Found 2 hits in 2 annotations

Action: Show Concordance View Page: << < > >> Hit 1 - 2 of 2 hits

Context Size: 4 Font: Arial Unicode MS 12 ☐ Show Info Balloons

so from **here.** yeah ja ja there is another plein

so from here. yeah ja ja **there** is another plein rotunda ja

Firefox IMDI Browser Trova Search Application MPI - ANNEX Interface

lat.csc.fi/ds/imdi_browser/ Google

IMDI-Browser show accessibility of resources about manual register user: anonymous login logout

Latserver

- WelcomeToLat.html
- Academic use
- Demo**
- Private
- info.html
 - a
 - c
 - e
 - g
 - h
 - i
 - j
 - k
 - l
 - lennes
 - liisamar
 - liuska
 - ljalava
 - Imkhuota
 - lttyys
 - luforsma
 - lxsalmei
 - m
 - n
 - o
 - p
 - r
 - s
 - t
 - v
 - w
 - z
- Public
- Restricted use
- Testing

IMDI

ISLE Metadata Initiative

Corpus

Name Demo	
Title	
Description	
This is a corpus for demonstration purposes	

DISTRIBUTING RESOURCES AND TECHNOLOGY

Distribution

- Infrastructures
 - CLARIN for research: www.clarin.eu/vlo
 - META also for industry: www.meta-share.eu
- Goal
 - Catalogue and make visible (“standardized metadata”)
 - Inform about usage conditions (“standard licenses”)
 - Promote interoperability (“plug-and-play”)

Funding Principles

- “Everyone should be able to generate taxable revenue streams on common goods”
- Publicly funded resources are made publicly available
 - EU currently applying this to government-created resources
 - USA also applies this to government-funded resources at private companies

ADDITIONAL INFORMATION

References

www.meta-net.eu/sra/

Strategic Research Agenda

ec.europa.eu/research/horizon2020/

Horizon 2020

Legal Exceptions

- **Exceptions**
 - Intellectual Property
 - Exception for “fair use” in USA, cf. Watson and Jeopardy
 - Exceptions also in the UK, Netherlands, Estonia, ...
 - Personal Data
 - Exception for research
 - Exception for distribution via anonymization and aggregates