Translate
Μία γλώσσα δεν είναι ποτέ αρκετή
Un solo idioma nunca es suficiente
Egy nyelv sosem elég
Una sola lingua non è mai abbastanza
Eine Sprache ist niemals genug
Знати један језик није довољно
One LANGUAGE IS NEVER enough
Dejjem tajjeb li tkun taf iżjed minn ilsien wiehed
Ett språk är aldrig nog
Yksi kieli ei koskaan riitä
Ní leor teanga amháin

Language Technologies in Context

LT-OBSERVE Brussels, November 2016

LANGUAGE TECHNOLOGIES IN CONTEXT

We are living in a connected world. Digital technology allows for seamless, ubiquitous communication that brings the world closer together than ever. But:

> "Like an ocean that unites continents and separates them at the same time, language is at the same time a bridge and a barrier between human beings". *Fritz Mauthner**

Without English as a vehicular language, communication is often impossible. Language is often a *de facto* barrier to the full deployment of political or economic goals like the Digital Single Market - by nature a multilingual market. "All languages have the same dignity" said Jean-Claude Juncker, President of the European Commission. Language technologies are the enabler for "language-neutral" content across Europe that can be accessed and understood by everybody in their own languages.

Like all digital technologies, language technologies are "vehicles" for reaching an ultimate goal. By their nature, LTs are closely related to content in various written and spoken languages and formats, and communication in a multitude of situations. ***Language Technologies can provide solutions to multilingual needs in relation to economy and societal challenges.***

*Fritz Mauthner, 1849-1923,Austro-Hungarian novelist, critics and philosopher who wrote "*Contributions to a Critique of Language" ("Beiträge zu einer Kritik der Sprache")* that fascinated Joyce and influenced Beckett.

TABLE OF CONTENTS

1 WHAT ARE LANGUAGE TECHNOLOGIES?

Language Technologies are digital technologies that support processes involving human languages, therefore often called NLP – Natural Language Processing. They are manifold but can be, roughly, divided into three main areas:

TEXT (TRANSLATION): From simple spell check to almost fully-automated "language transfer"

SPEECH (INTERACTION): From spoken human-machine interaction to speech synthesis

ANALYTICS: analysis of text/speech data to gather deep insights, often complemented by sentiment analysis

TEXT (TRANSLATION)

Language Technology tools for written language range from simple spell-checkers to computer-aided translation (CAT). What is commonly known as machine-translation is usually an automated language transfer process whereby the machine substitutes phrases from a source into target language without the intellectual effort that full translation implies. Translation between two languages is by far the most complex and rich application of textual technology at present.

Machine translation can use *statistical models* based on automatically predicting equivalent phrases between two languages from a large amount of ***parallel corpora***: A parallel corpus contains a collection of existing translations, with the original texts in one language and their aligned translations in another language. Statistical machine translation systems can be developed relatively quickly if the parallel data exist to spot "parallels" between source and target languages.

Rule-based machine translation, on the other hand, is based on linguistic information (lexical, grammar), morphological, syntactic, and semantic analysis derived from monolingual and/or bi- or multilingual sources. The rules have to be hand-crafted, so it takes much effort, cost and time to develop rule-based systems.

So far, neither system yields qualitatively satisfactory results for all language pairs. It is likely that hybrid systems (part statistical, part rule-based) or the currently emerging *neural machine translation*[1] will drive the future of machine translation.

Currently available examples[2]:

- ➔ Not only Google: SDL Free Translator: https://www.freetranslation.com/
- ➔ EC – DG Translate Acquis Communautaire and translation memory based on it, covering 552 language pairs.

  https://ec.europa.eu/jrc/en/language-technologies/dgt-acquis

  https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
- ➔ A Latvian – English and Latvian - Russian machine translation tool created on the occasion of the Latvian EU presidency. HUGO https://hugo.lv/en
- ➔ Linguatec Personal Translator http://www.linguatec.net/products/tr/pt : ready-to-buy product with clear price indications for private and business use packages.

SPEECH (INTERACTION)

Speech technology provides two forms of language technology – speech recognition and text-to-speech. Speech recognition transforms spoken language into a text output (like dictation) but also enables the natural speech interaction of humans with machines (e.g. spoken orders into mobile telephones).

Currently available examples:

[1] For a definition, see: http://104.131.78.120/ (LISA, University of Montreal); neural MT uses recurrent neural networks to map input sequence (source language) with output sequence (target language). See also: https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

[2] All examples given are provided as illustrations and do not present preferred solutions or advertising

- ➔ *C*ar speech dialogue systems: to control radio, telephone, TV tuner and main functions of GPS with voice
- ➔ Automatic translators for travellers**:** itranslatevoice.com App for mobile telephones speech to speech
- ➔ Automatic subtitling and/or caption: not yet perfect, YouTube and Watson (IBM) make some efforts.

(DATA) ANALYTICS

Data Analytics uses algorithms to automatically "read" raw text data to gather insights and draw meaningful conclusions from the contents. It is also referred to as a "business intelligence" tool when applied in a commercial or entrepreneurial context because it helps enterprises and organisations organise and optimise their search, understanding and leverage of textual content. Data can include (multilingual) text data but also figures, multimedia content and sensor data.

Currently available examples:

- ➔ Sentiment analysis in social media texts: http://sentistrength.wlv.ac.uk/ some free software *on-line available; licence fee for commercial purpose 1000 Pounds.*
- ➔ Media mining and analysis**:** http://www.sail-labs.com/products-solutions.html

COMBINED TECHNOLOGIES

Language Technologies are often **embedded in other technologies** or combined to produce richer services. Machine translation, for example, can be combined with Analytics to make sense of and draw conclusions from multilingual data; or with Speech, to create speech-to-text subtitles.

Content management systems use APIs to provide a multilingual layer by drawing on rich language tools and resources that provide the resources.

Other emerging technologies such as **augmented reality** enable telephones to take a photo of a name in an unknown alphabet and have the underlying software recognise the language and translate it into the user's tongue.

Automated news story generation uses language technologies with **Artificial Intelligence** algorithm to produce routine news stories based on clean, structured data such as sports reporting or finance reports, without needing human intervention to do the writing. Artificial Intelligence is also at the root of (sentiment) analytics, or neural MT.

2 **WHY** LANGUAGE TECHNOLOGIES?

> "The language of Europe is translation"
>
> *Umberto Eco, 1993*

Multilingualism is a cultural asset of Europe – but also a barrier. While English acts as a *lingua franca* for international communication, only 31 % of internet users speak English as their native language. Social media are also rapidly becoming multilingual. In 2015, more than half of the 200 million tweets per day were written in a language other than English. Eurobarometer reported that nine out of 10 European internet users prefer to browse in their native language, and 70% of European on-line buyers make purchases in their native language only.

The EU has **24 official languages** making no fewer than **552(!) language pairs**. This ignores all the unofficial but widely used local languages that add up to more than 60 languages in the EU alone.

Creating a seamless **Digital Single Market,** therefore, is even linguistically not a trivial task.

By nurturing a Digital Single Market, Europe can create up to €250 billion in additional growth, hundreds of thousands of new jobs, and a vibrant knowledge-based society. This is an enormous opportunity that Europe cannot afford to lose. But it will only work effectively if the language barrier (as well as the others) is overcome effectively.

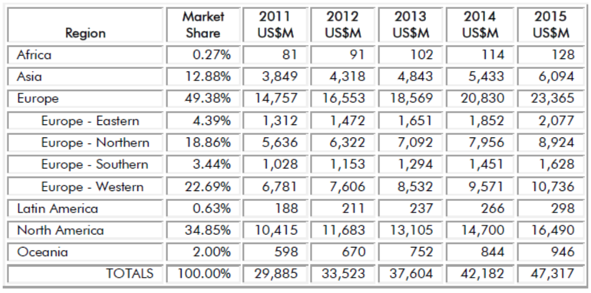
ECONOMIC IMPACT

1. On Language Technology Providers

The fast-growing Language Technology (LT) industry includes all three strands mentioned above (speech, text and analytics). It therefore includes the widely known Languages Services Providers (LSP) industry, focused on translation, interpretation and localisation, using a mix of human and technology resources.

> ”Europe is a natural ecosystem for language technology”
> *Jochen Hummel, Chairman, LT-Innovate*

Common Sense Advisory, Inc. has regularly surveyed the LSP sector per region for five years:



Region	Market Share	2011 US$M	2012 US$M	2013 US$M	2014 US$M	2015 US$M
Africa	0.27%	81	91	102	114	128
Asia	12.88%	3,849	4,318	4,843	5,433	6,094
Europe	49.38%	14,757	16,553	18,569	20,830	23,365
Europe - Eastern	4.39%	1,312	1,472	1,651	1,852	2,077
Europe - Northern	18.86%	5,636	6,322	7,092	7,956	8,924
Europe - Southern	3.44%	1,028	1,153	1,294	1,451	1,628
Europe - Western	22.69%	6,781	7,606	8,532	9,571	10,736
Latin America	0.63%	188	211	237	266	298
North America	34.85%	10,415	11,683	13,105	14,700	16,490
Oceania	2.00%	598	670	752	844	946
TOTALS	100.00%	29,885	33,523	37,604	42,182	47,317

Table 1: Language Services Market Share by Region

The results show a steady increase in all regions, and most significantly in Europe. However, US giants such as Google and Microsoft continue to invest deeply in LT. India too has decided to invest in knowledge resources to overcome its fragmented language landscape with 1000+ languages (http://www.newsgram.com/technology-to-bring-indian-languages-together/) while EU investments in LT are now stalling. The Americans are investing particularly in technology methods such as machine learning and Artificial Intelligence to transition to the next stage of digital development as a whole, in which language processing (LT) contribute to a broad range of applications, from personal digital assistants to improved speech recognition technology to social robotics. For Europe and India, the main technology issue is how to overcome the translation barrier at a time when human resources and workflows are insufficient to meet the needs of a fast, connected world.

2. Economic impact on businesses benefitting from LT:

The Digital Single Market can only be realised if cross-border commerce is truly seamless – i.e. every European citizen can sell or buy wherever s/he wants in any language s/he wants to use, without geo-blocking or a translation failure. Without this, Europe will still remain a mega-market of many fragmented local markets that are not big enough to compete globally with big players.

SOCIAL AND SOCIETAL IMPACT

We talk a lot about mobility and inclusion, and most recently, we have been faced with such new challenges as refugees and the threat of terrorism. Language technologies are not a panacea, but they can support society at many levels – provided that they are used to create ”language-neutral” content that concerns the most urgent demands of society and government:

Product information on allergenic ingredients for comestibles

Patient information for travellers with diseases

Human-machine voice interaction for the visually impaired or elderly with reduced mobility

Patient–doctor interaction for travel or immigrant situations

Intuitive language learning using audio-visual modules

Data analytics (including social media analytics) for predicting trends (refugee waves) or for preventing disasters (human-made or natural)

3 **WHAT** IS MISSING?

LANGUAGE RESOURCES

Both, text and speech technology depend for their effectiveness on accessing and learning from large quantities of existing data. These "language resources" form the database for monolingual and parallel corpora that create a "translation memory" for machine translation or speech resources that can be used to train speech technology. This forms part of a much larger and global technology trend towards machine learning based on massive data input.

Today, only English has a sufficient amount of such language resources to train technology systems effectively. All other languages lag behind. This is due to:

- ➔ **restricted use** of existing resources (e.g. they can only be used for research, not commercial applications), copyrighted data or data restricted by privacy and trade-secret laws;
- ➔ **minimal resources** (the cost and time is prohibitive) are available for less used languages in the EU

NOT YET PERFECT LT TOOLS

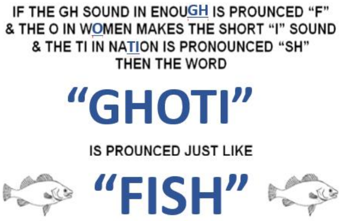
Because **machine-assisted translation** does not have the language intelligence of a human being, there are other obstacles (apart from the above-mentioned lack of resources) that prevent fully-acceptable translation quality in LT. Amongst them are **disambiguation** and **proprietary names**.

To give a real example:

> When testing the HUGO system (EN-Latvian), I typed in the phrase: "Who can use HUGO?" the answer in Latvian was: "Kurš var izmantot Igo?" Although HUGO is the proprietary name of the system, the system itself translated it into the Latvian form of the name, i.e. Igo.

Speech recognition is also a major challenge due to the fine distinctions in **human phonetic particularities**. One – not too serious - illustrative example is given by

https://docs.google.com/presentation/d/1PU-hn68nZcItDASLQ_1g43iEp6fY6tHMvujSME5y03U/edit#slide=id.g1dc5f4f7e_235



IF THE GH SOUND IN ENOUGH IS PROUNCED "F"
& THE O IN WOMEN MAKES THE SHORT "I" SOUND
& THE TI IN NATION IS PRONOUNCED "SH"
THEN THE WORD

"GHOTI"

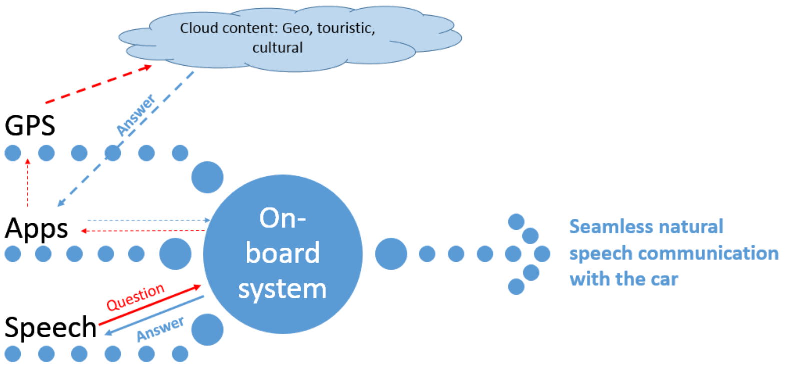
IS PROUNCED JUST LIKE

"FISH"

COMPREHENSIVE LT ECOSYSTEMS

Despite the technical challenges, Europe has all the ingredients needed to make an LT ecosystem work. The three critical weak points are: **Interface – Interoperability – Integration.**

Here's an example from the ***automotive sector*** (simplified):


Cloud content: Geo, touristic, cultural
GPS
Answer
Apps
Speech
Question
Answer
On-board system
Seamless natural speech communication with the car

A driver or passenger could ask the system any of the following questions:

- How is the weather at my destination?
- Is there any cheap gas station on the route?
- What's that high red building over there / on my left?
- What's the speed limit here?
- How far am I from the neighbouring car?

To provide answers to these questions, several components must be activated, integrated and rendered interoperable:

- On-board and on-line communication tools (text, speech in all available languages for drivers)
- Language-independent ontology (common knowledge categories underlying the system' understanding module)
- Geopositioning
- Sensor information
- Apps (e.g. booking of tickets or rooms)

Currently, many single modules of this overall package are available (e.g. voice commands; sensor alerts) but the full cloud-car ecosystem with interoperable plug-ins is not yet a reality.

Similar missing links can be observed in other vertical value chains, such as media and publishing, (language) learning, pharma/chemicals, health care, and more. The Suggested Actions below give an indication as to how the most pressing developments can be realised with the help of public players.

4 **SUGGESTED ACTIONS** AT EU, NATIONAL AND REGIONAL LEVEL

OVERALL: GENERAL AWARENESS NEEDS

- Awareness-raising among politicians, decision-makers, and funding agencies as to what LT is all about
- Awareness-raising of the benefits that LT can bring at economic, social and societal levels

CREATE AN LT-SUPPORTING IPR REGIME

In order to make machine translation (MT) work, an MT engine needs to "learn" the languages it is fed with. Statistically, this is done by processing a large amount of parallel corpora (2 or more languages) so that the engine learns on an "if…then" principle: If a sentence in language X looks like this, then it will probably look like that in language Y, based on millions of examples in the corpora.

For rule-based machine translation (or hybrid solutions), monolingual resources can be used to smooth the quality of the target language output, again based on statistical facts about the morphological, syntactic, and semantic structure of the language. What this boils down to is that large amounts of (relevant) content are needed to improve the tools that perform these kinds of operations. This is a special challenge for less-used languages for obvious reasons – there is simply far less digital data available.

Opening up public data for machine-learning purpose would be a first step (always considering all-purpose-use, including commercial use!). An exemption from copyright is still only wishful thinking but would be a remedy in particular for languages with few Language Resources.

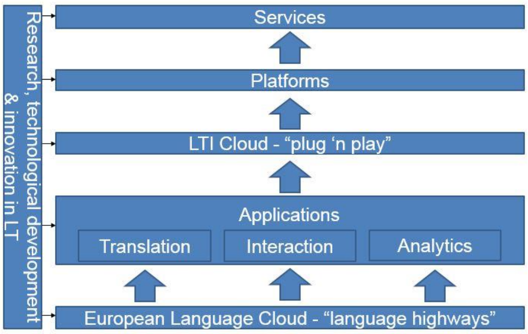
CREATE AN EU LT STRATEGY

A borderless infrastructure is needed to guarantee access to all (official) languages and tools for all European citizens, businesses, researchers, and public administrations. Only language-neutral content, be it cultural, business or public, can make the Digital Single Market genuinely "single".

> "Language barriers are market barriers", *Georg Rehm, META-NET*

The vision of an ideal "LT ecosystem" would therefore look like this:

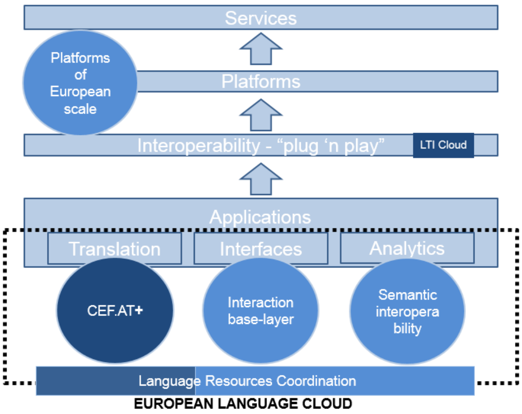
Proposed LT strategy for the EU in a nutshell


Research, technological development & innovation in LT
Services
Platforms
LTI Cloud - "plug 'n play"
Applications
Translation
Interaction
Analytics
European Language Cloud - "language highways"

In this graph, the European Language Cloud layer forms the foundation for all other tools, products and services. We therefore call it "language highways" as they are needed to "transport" language resources in a connected, open, borderless way to all the destinations required by innovation and ultimately the market.

Currently, bits and pieces exist or are being implemented, yet the overall strategy is not yet fully recognised. This means that there are inevitable gaps, and those pieces that do exist tend to underperform.

The second figure shows where action at European has started and where it is still needed:



The dark blue elements are currently being implemented. These include the Automated Translation (AT) for the Connecting Europe Facility intended to make all Digital Service Infrastructure multilingual (but currently restricted to use by the public sector), and the LTI Cloud that federates tools from the LT businesses across the EU in a common cloud marketplace to provide easy access, especially for SMEs. At the infrastructure level, there are two urgent action areas - speech technology and semantic interoperability. There are plans from successfully completed EU projects (e.g. CITIA for speech technology), but the creation of a Europe-wide infrastructure needs initial financial support at EU level to avoid immediate market fragmentation into language or sector.

At the platform level, much can be done by the private sector or through research and innovation projects. However, truly European scale platforms are needed to implement the Digital Single Market. This is not only a question of strategy, but of concerted action and financing.

SUPPORT LT AT NATIONAL/REGIONAL LEVEL

Europe has a highly diversified language landscape which adds to the cultural richness of Europe but raises *de facto* barriers for a single market.

There are two crucial ways to support this drive to a multilingual digital singe market at national/regional level so that Europe remains multilingual and the digital economy is open for all:

-***Reuse of public sector content*** for language resources. Less-used languages in the EU suffer from machine translation efforts due to the lack of substantial digital language resources. New resources must be found or created, and public sector information is the most obvious place to search for relevant content to build such multilingual resources.

- Use of national/regional funding for innovative language projects for national or regional languages would provide an excellent source of support for this strategic endeavour. Results from any projects involving local/regional language data collection/creation could plug smoothly into the different layers of the ecosystem as described above.

IMPRINT

The LT in Context paper is a collaborative effort of the LT Observatory project. This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 644583.

Its partners are: LT-Innovate, the Association of Europe's LT Industry; EMF, the Forum of e-Excellence; CLARIN-ERIN – the Common Language Resource and Technology Infrastructure; UniVie – the Centre for Translation Technology of the University of Vienna, and Zabala Consultancy (coordinator, Spain).

Editor: Margaretha Mazura, EMF, mm[at]emfs.eu

DISCLAIMER:

All information was collected using the utmost diligence. However, the authors cannot be held liable for any inaccurate information or the use made of it. Neither can the European Commission be held responsible for any use which may be made of the information contained herein.