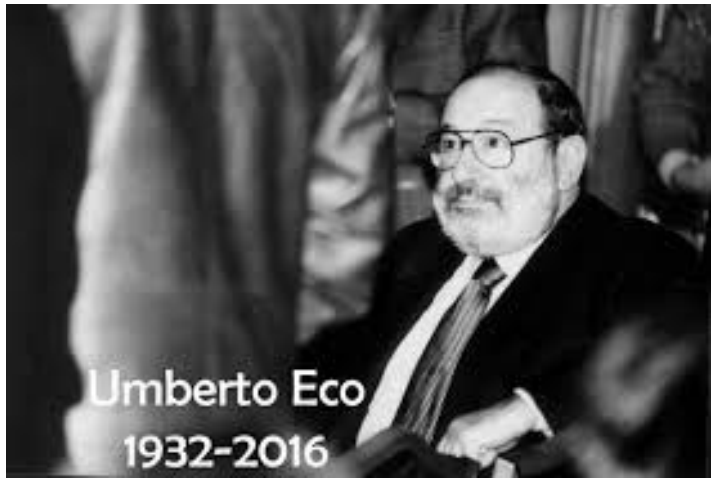


The language of Europe is translation.

« La langue de l'Europe, c'est la traduction. »

Assises de la traduction littéraire à Arles (France) le 14 novembre 1993,

Human (Multilingual) Language Technologies for a Multilingual Europe

Khalid CHOUKRI

ELRA/ELDA
choukri@elda.org

2019 | INTERNATIONAL YEAR OF
Indigenous Languages

https://en.unesco.org/news/unesco-launches-website-international-year-indigenous-languages-iyil2019

Draft prospective for a World Summit on

"Language Technologies for All (LT4All)"

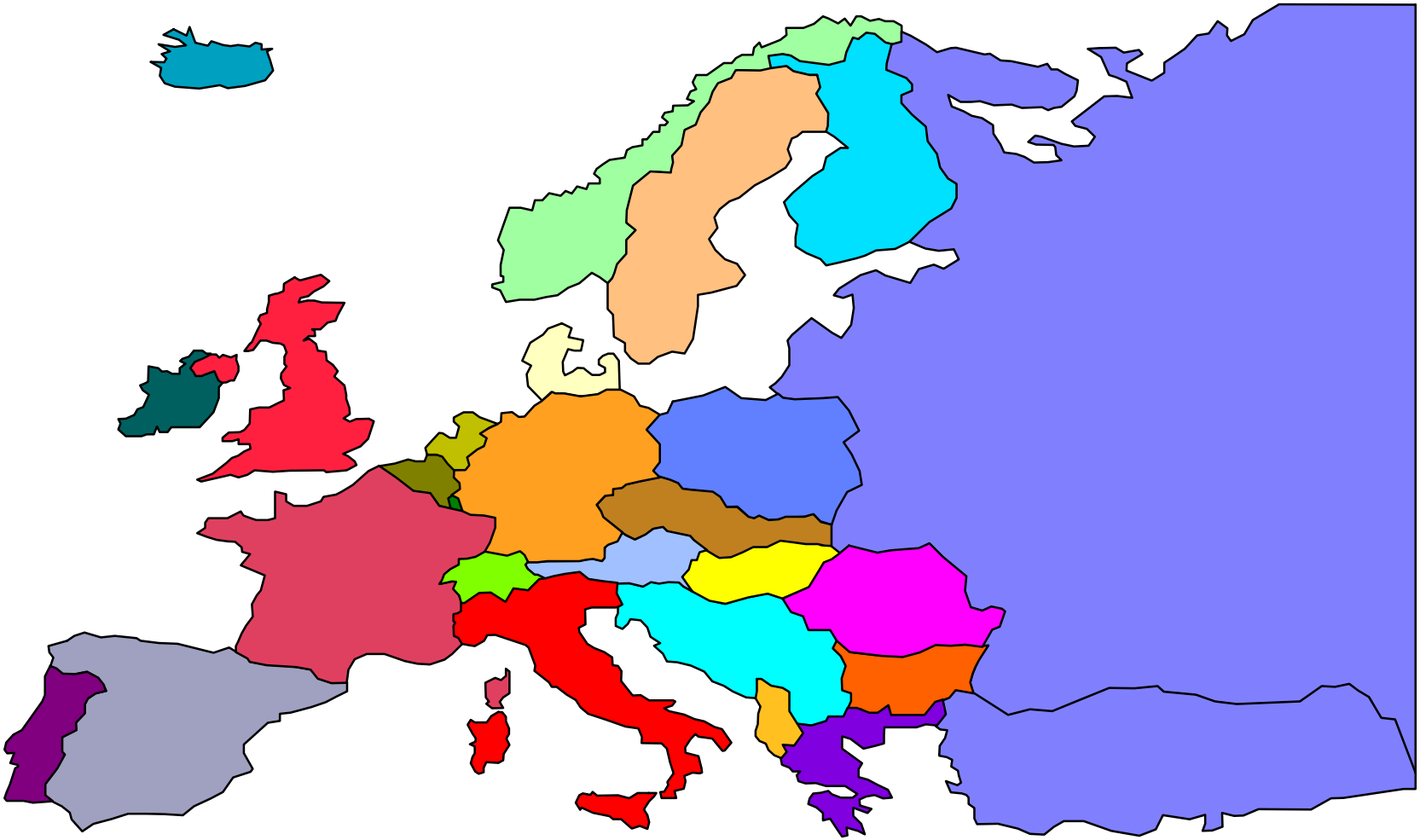
in the framework of the

UNESCO Year of the Indigenous Languages – 2019

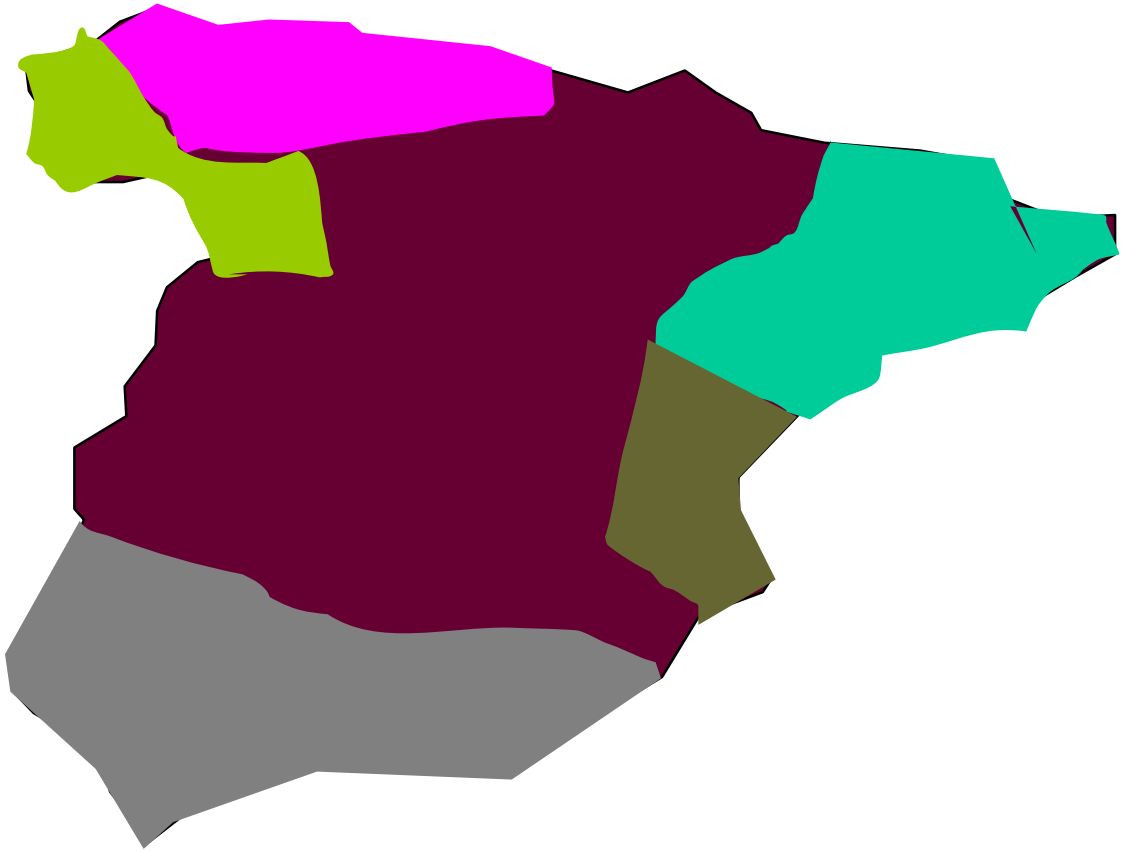
[ISCA/ELRA SIG-UL: Special Interest Group on Under-resourced Languages]

29/01/19 LT4All Meeting @ ELRA 1

Europe and the Multilingual Issue

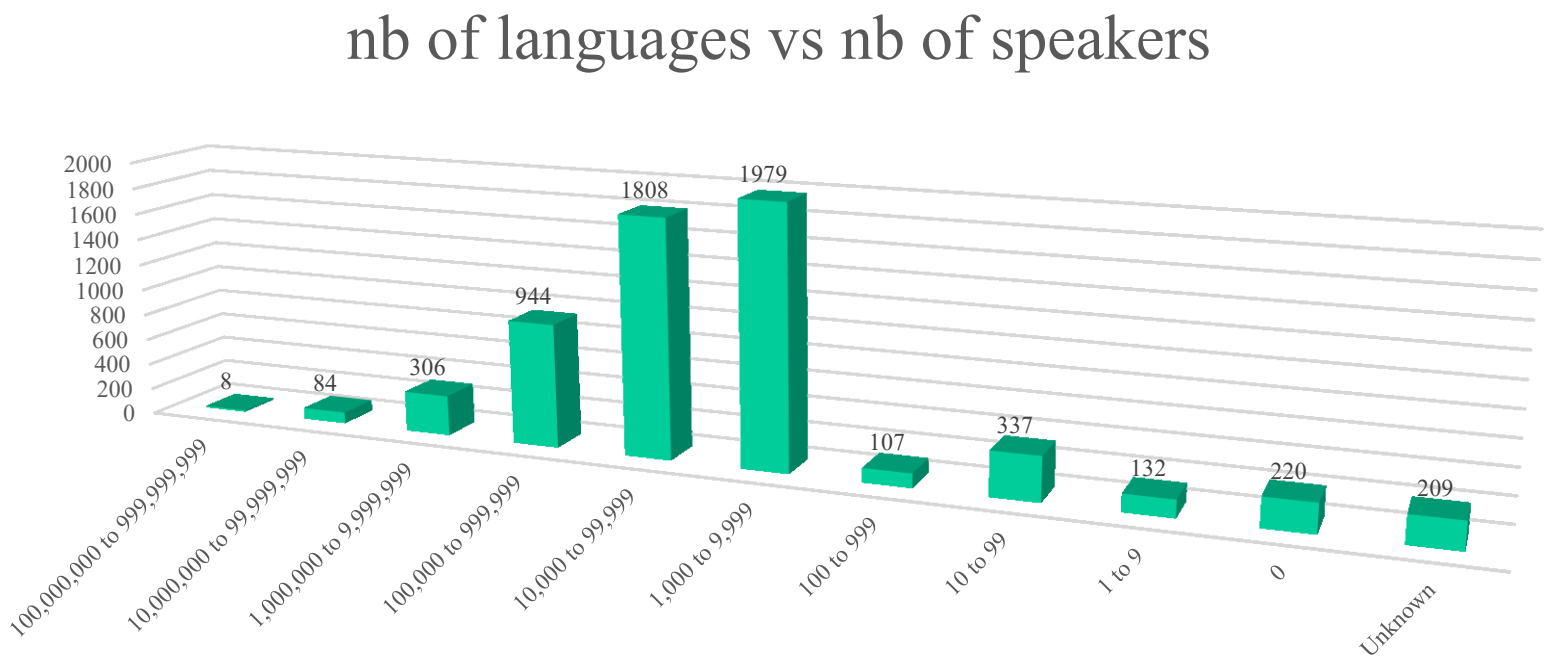


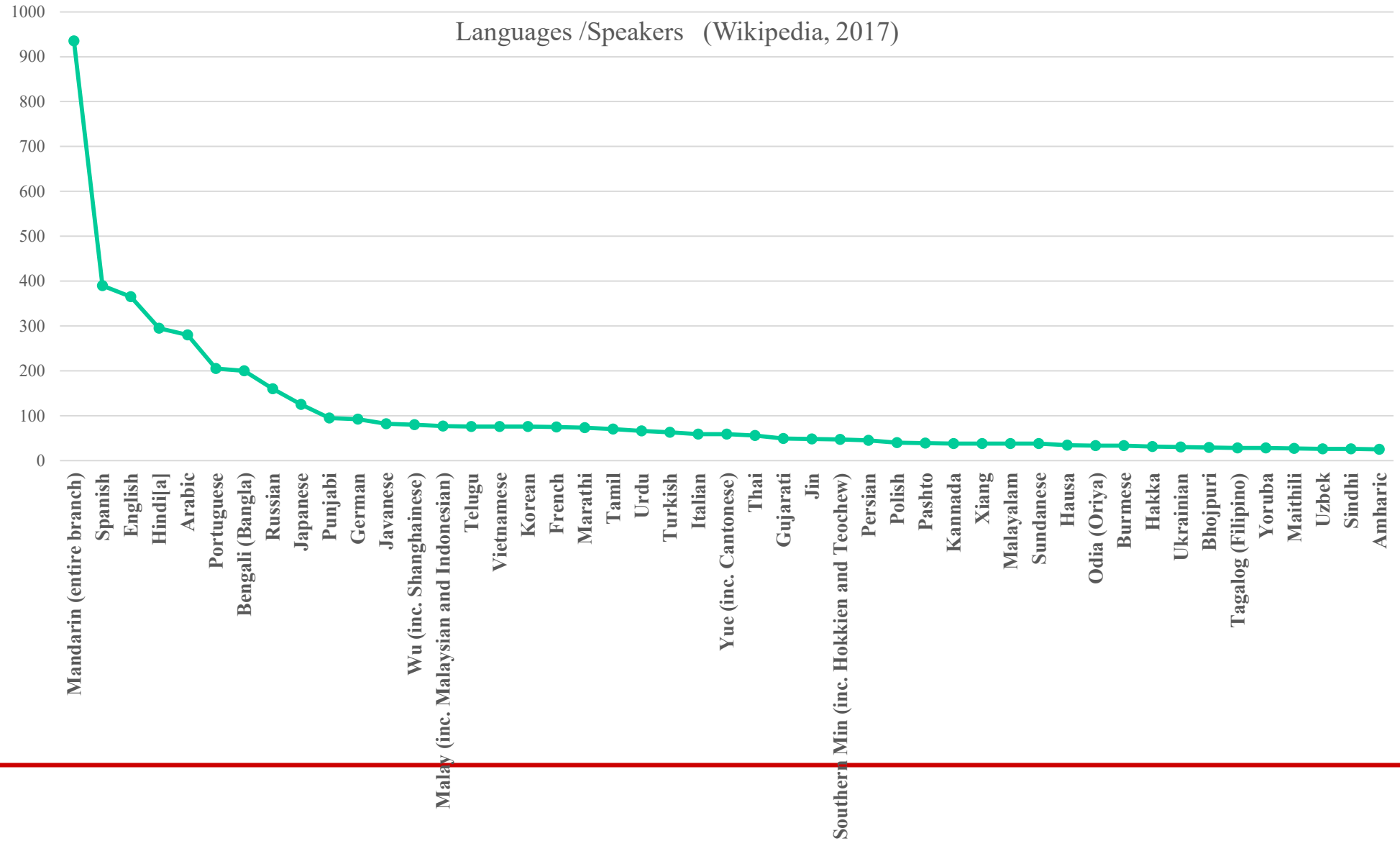
Even More Languages !



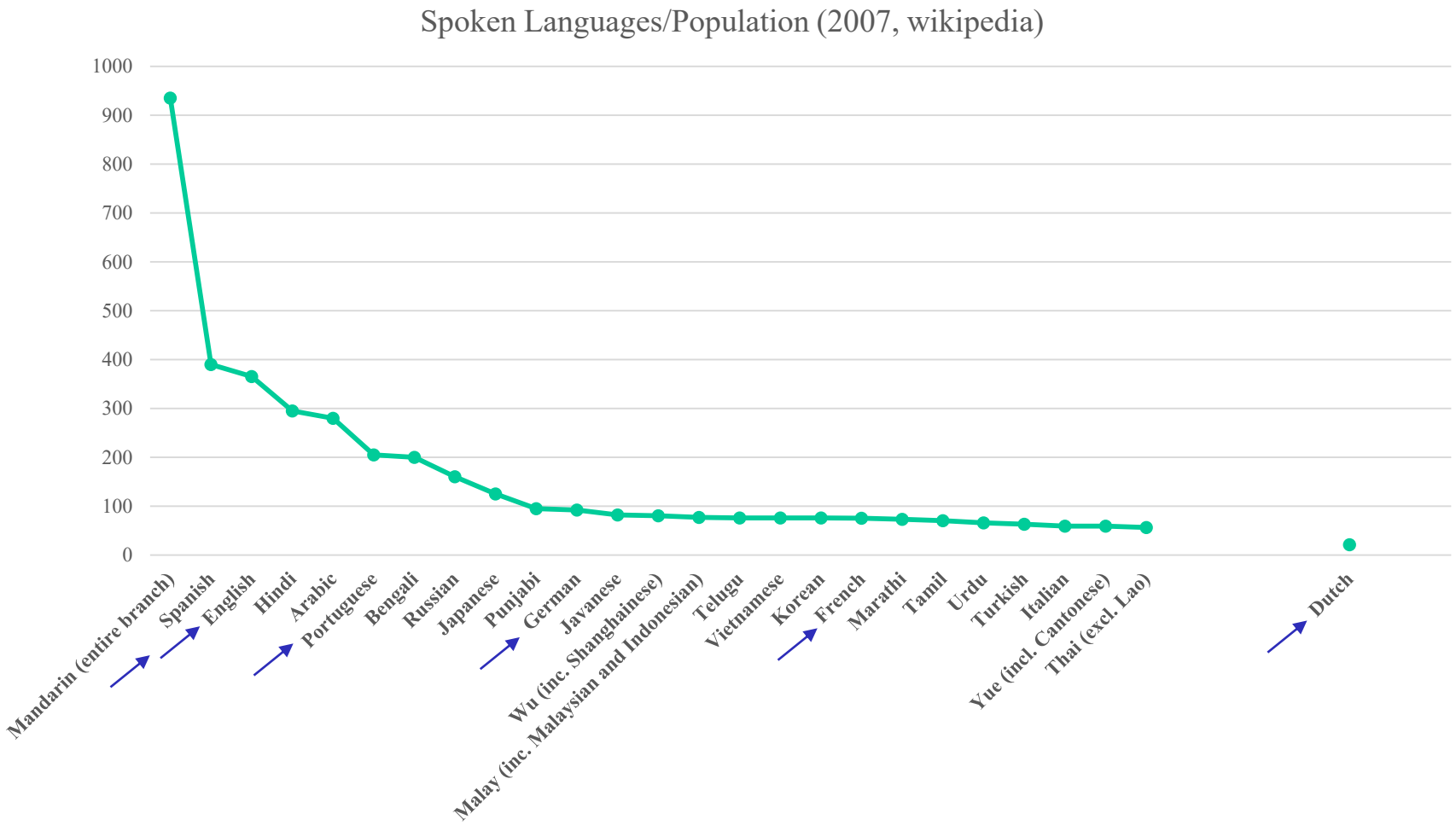
Languages: instruments of culture, identity and business

1. Over 7000 Languages
2. Languages have multiple modalities: **spoken**, written, signed
3. …. Only 200-300 have writing systems (about 50 different systems)

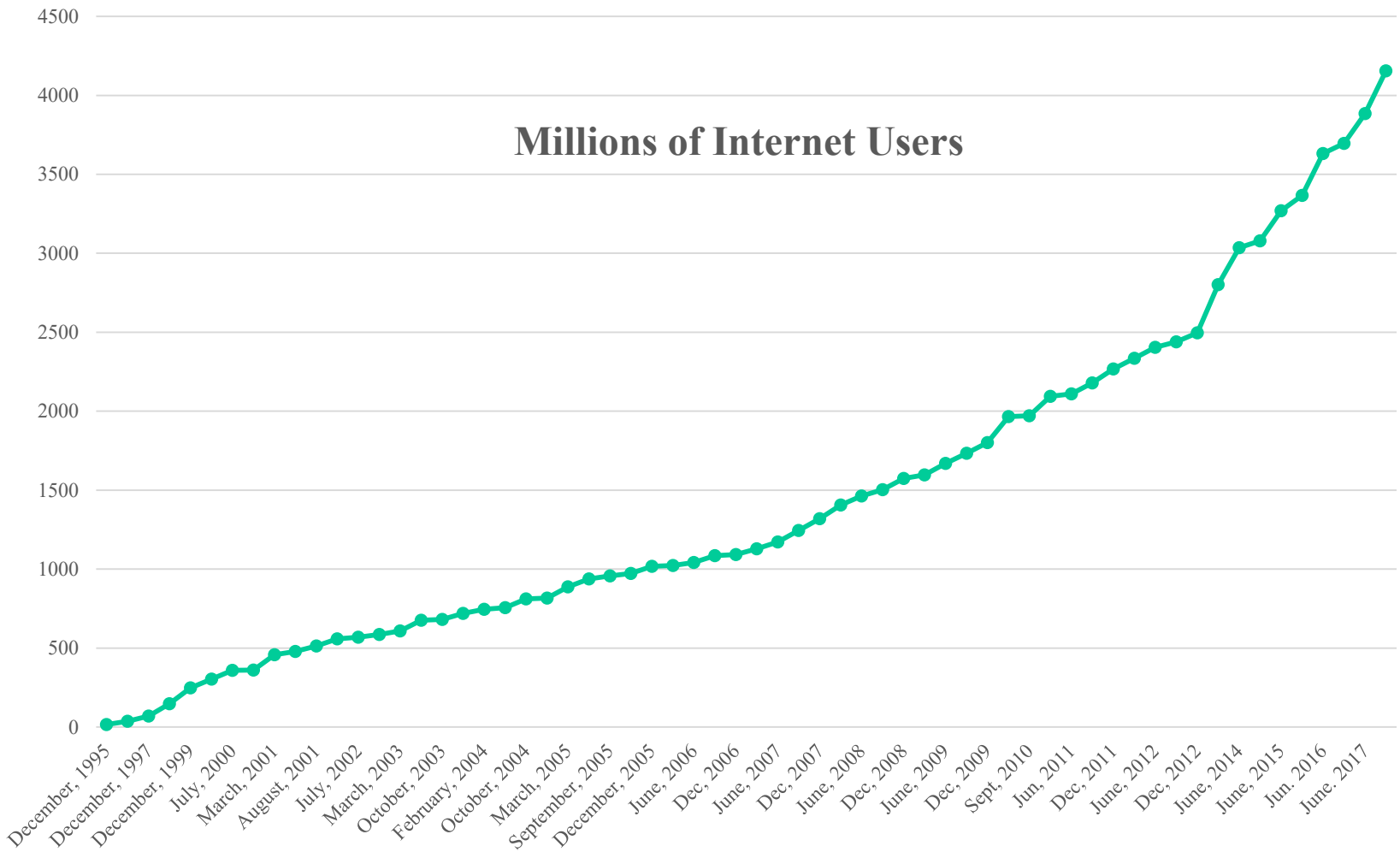



Languages /Speakers (Wikipedia, 2017)
1000
900
800
700
600
500
400
300
200
100
0
Mandarin (entire branch)
Spanish
English
Hindi[a]
Arabic
Portuguese
Bengali (Bangla)
Russian
Japanese
Punjabi
German
Javanese
Wu (inc. Shanghainese)
Malay (inc. Malaysian and Indonesian)
Telugu
Vietnamese
Korean
French
Marathi
Tamil
Urdu
Turkish
Italian
Yue (inc. Cantonese)
Thai
Gujarati
Jin
Southern Min (inc. Hokkien and Teochew)
Persian
Polish
Pashto
Kannada
Xiang
Malayalam
Sundanese
Hausa
Odia (Oriya)
Burmese
Hakka
Ukrainian
Bhojpuri
Tagalog (Filipino)
Yoruba
Maithili
Uzbek
Sindhi
Amharic

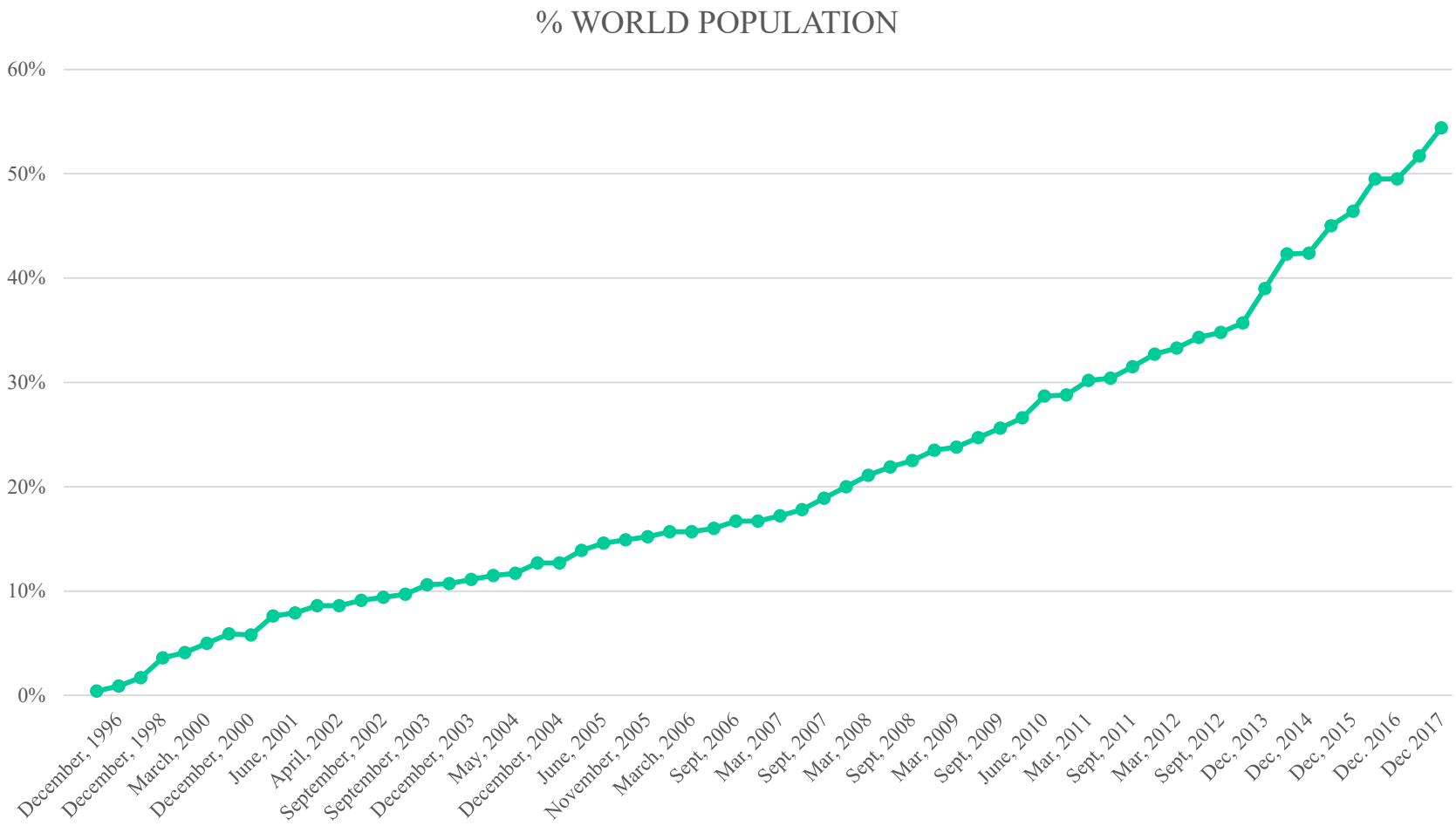
Languages: instruments of culture, identity and business



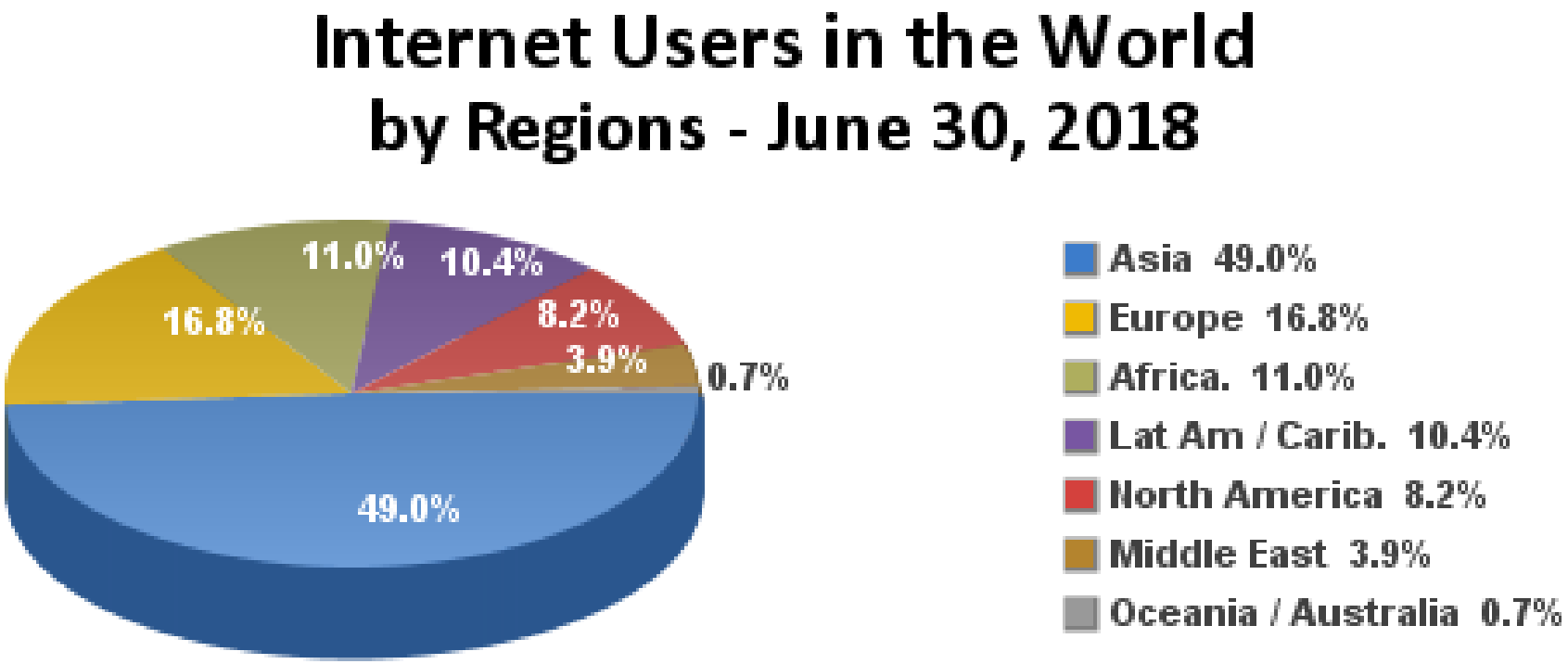
Internet Users over time



Ratio Users of Internet

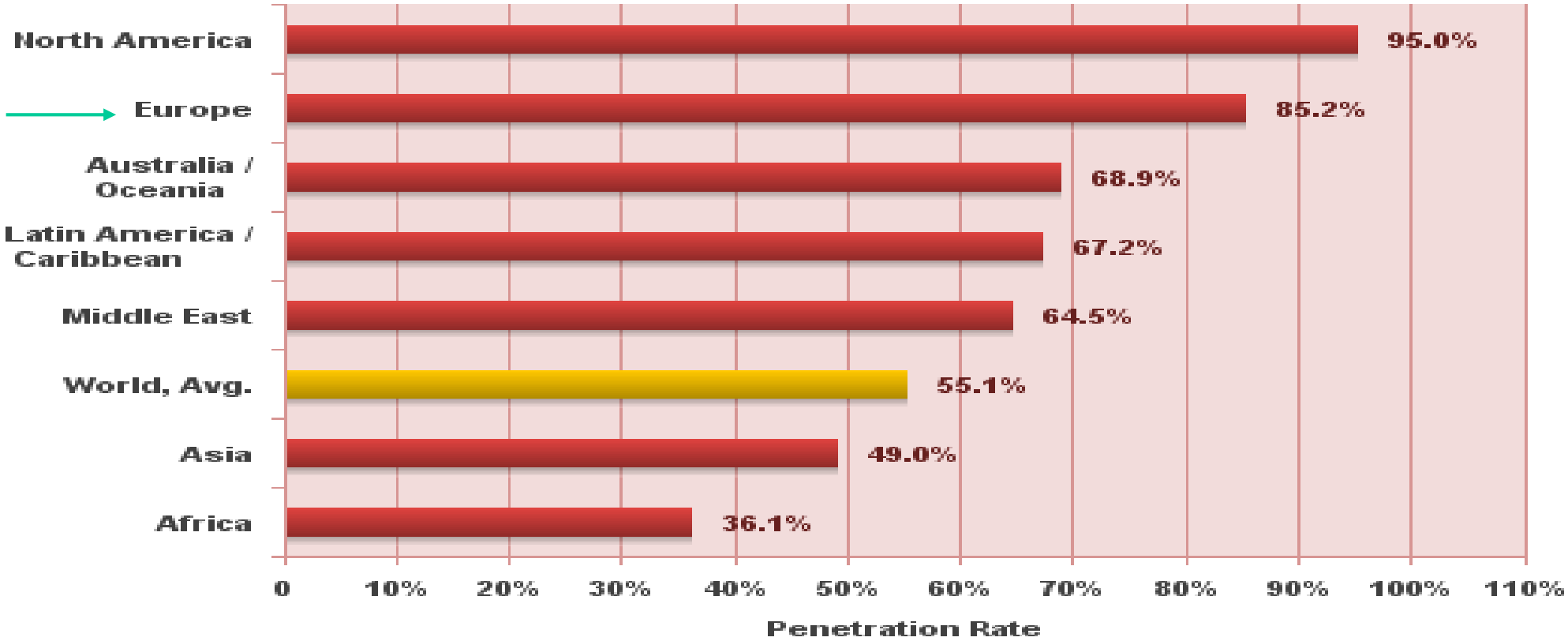


Cyberspace Evolution (mostly web)



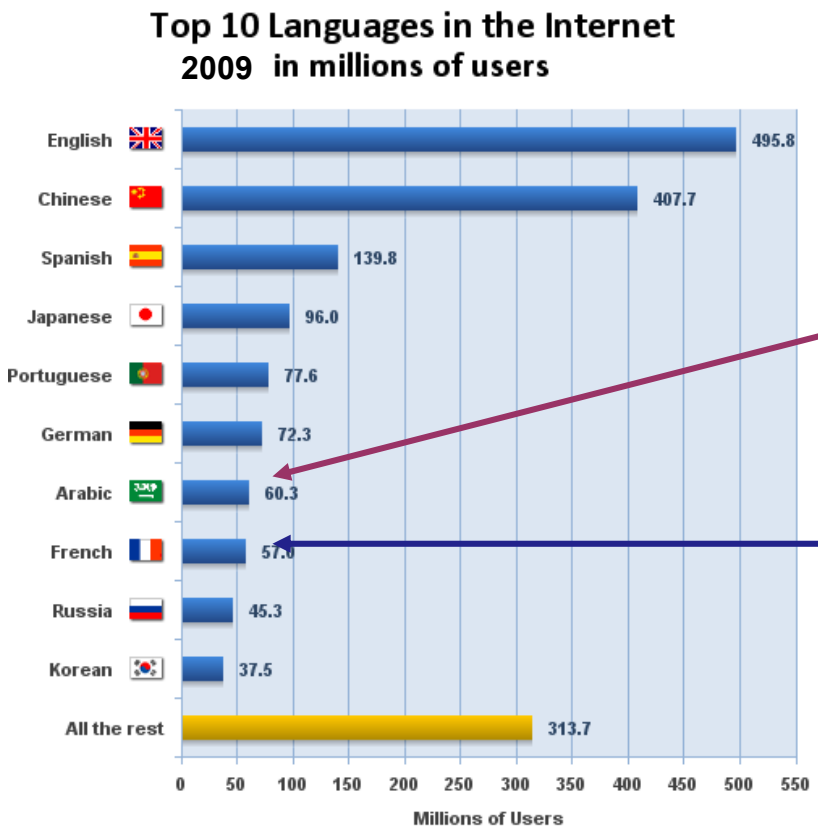
Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 4,208,571,287 Internet users in June 30, 2018

Internet World Penetration Rates by Geographic Regions - June 30, 2018

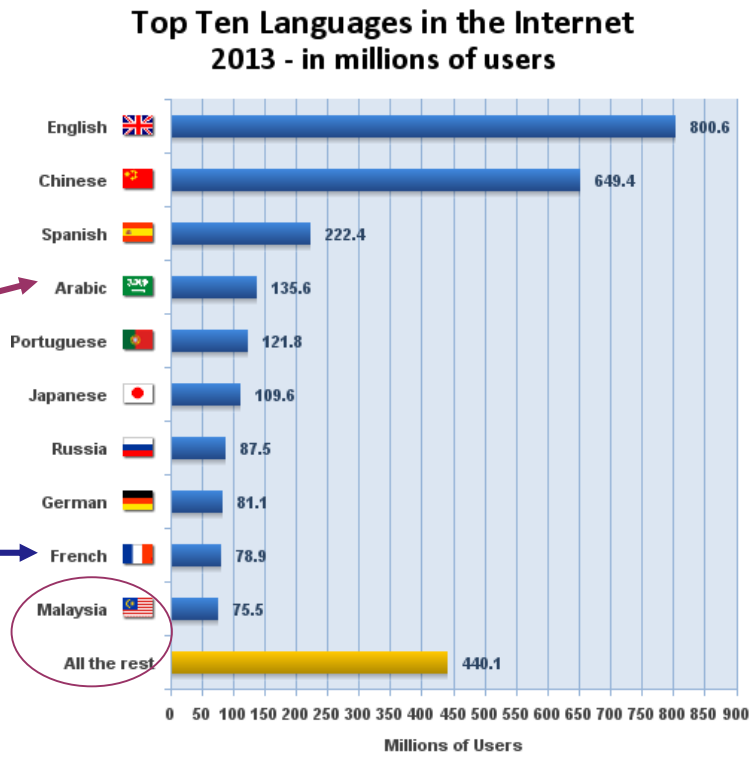


Source: Internet World Stats - www.internetworldstats.com/stats.htm
Penetration Rates are based on a world population of 7,634,758,428 and 4,208,571,287 estimated Internet users in June 30, 2018.

Langauge Cyberspace Evolution (mostly web)

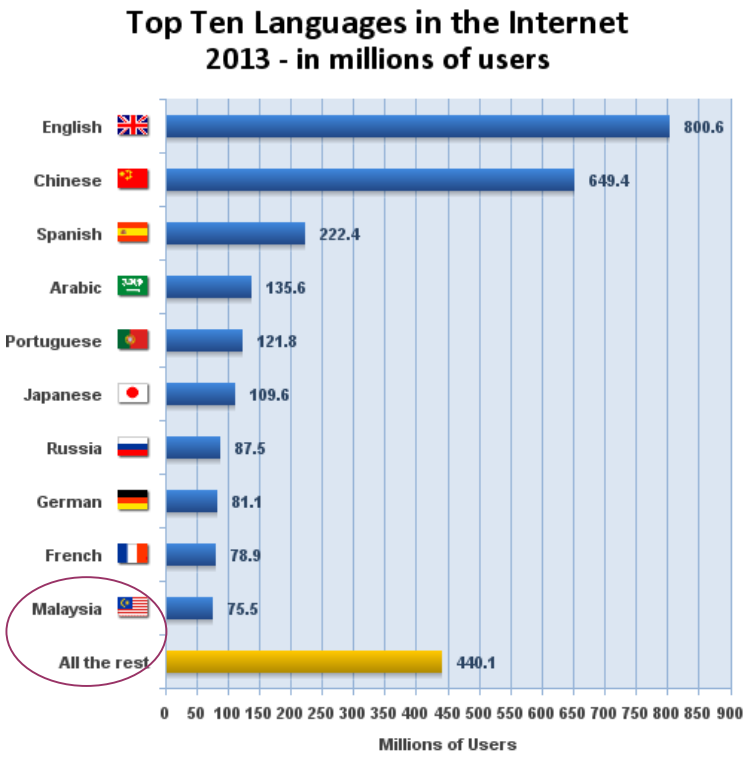


Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 1,802,330,457 for December 31, 2009
Copyright © 2000 - 2010, Miniwatts Marketing Group

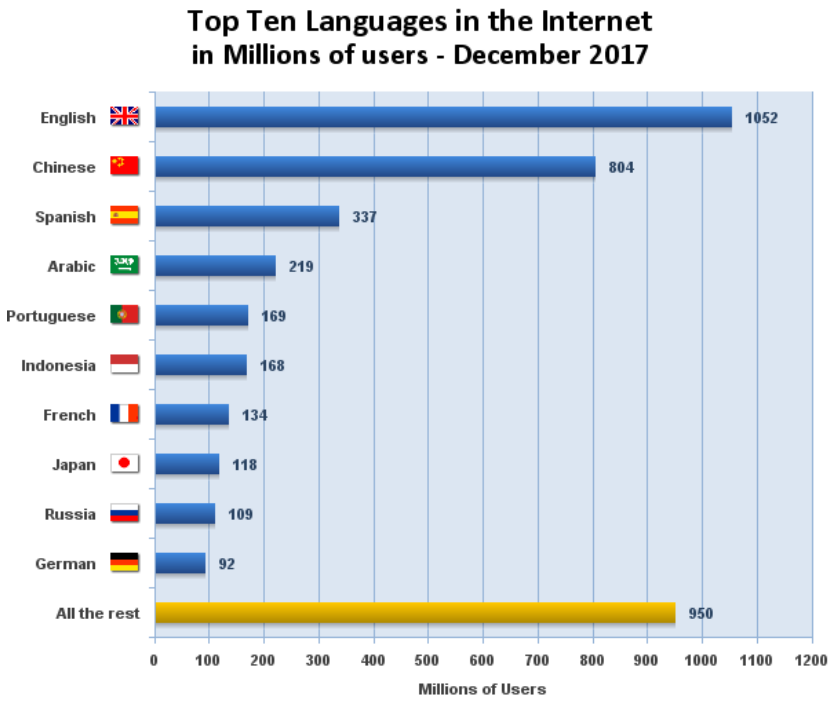


Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 2,802,478,934 on December 31, 2013
Copyright © 2014, Miniwatts Marketing Group

Cyberspace Evolution (mostly web)

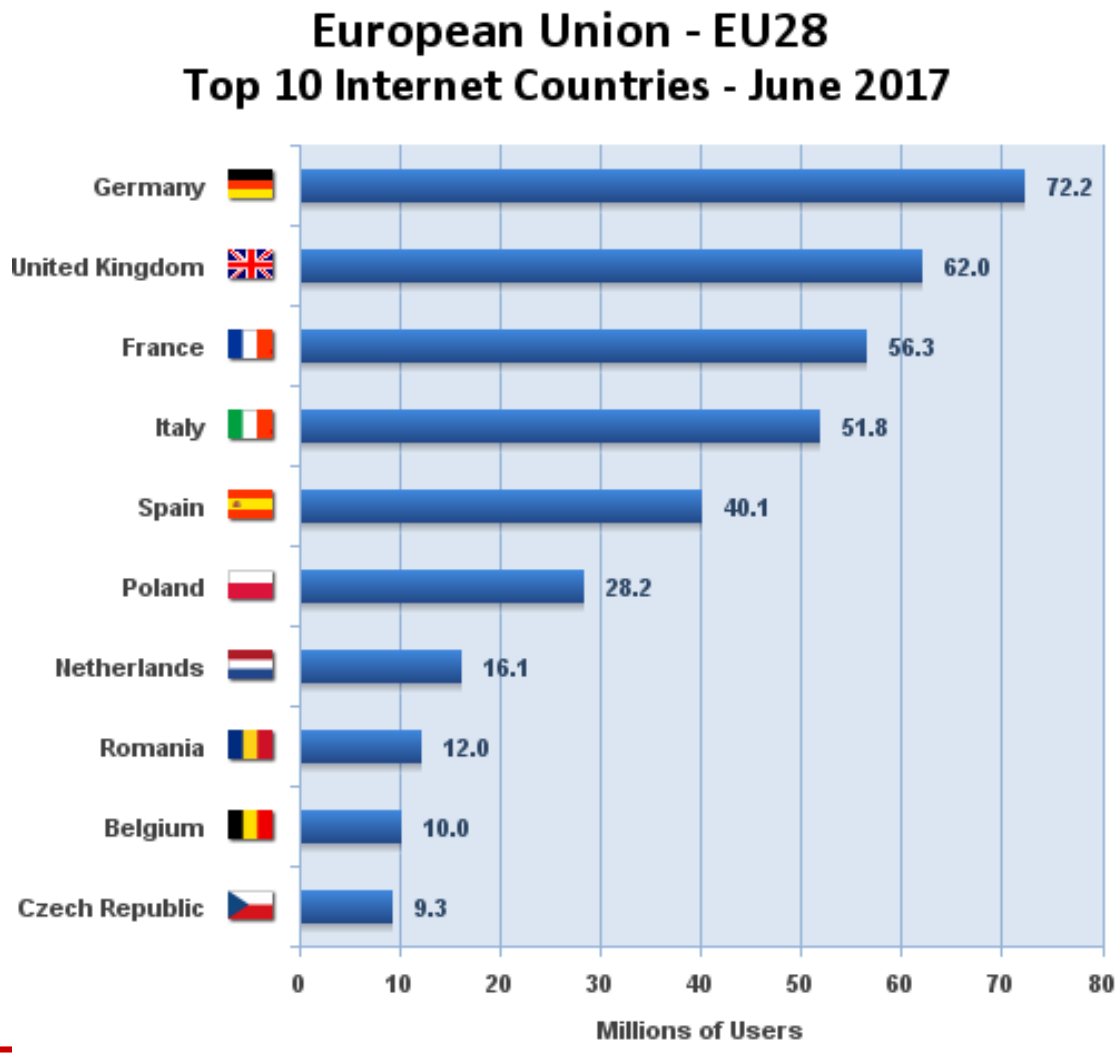


Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 2,802,478,934 on December 31, 2013
Copyright © 2014, Miniwatts Marketing Group



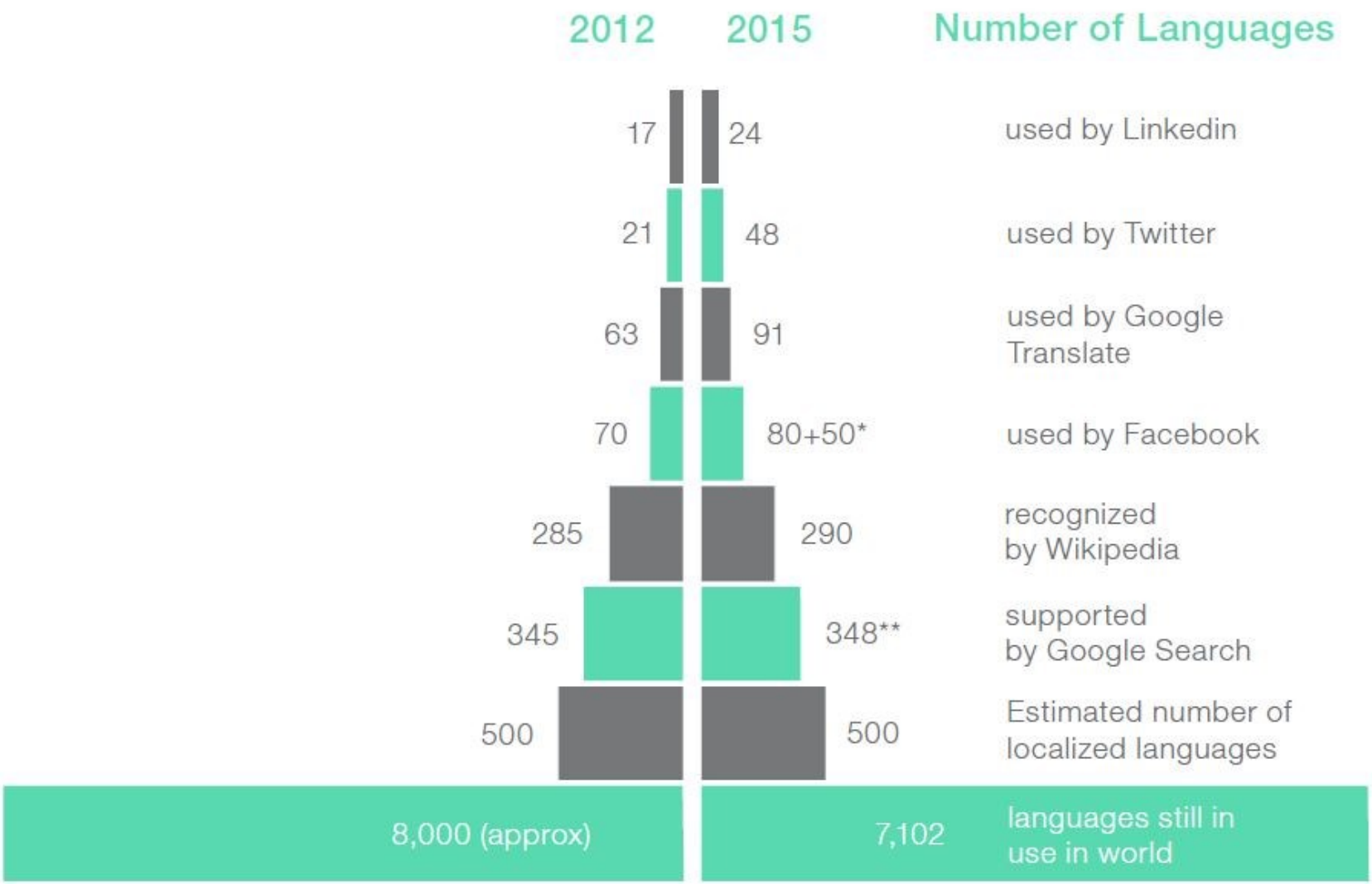
Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated total Internet users are 4,156,932,140 in December 31, 2017
Copyright © 2018, Miniwatts Marketing Group

At the EU Level ….

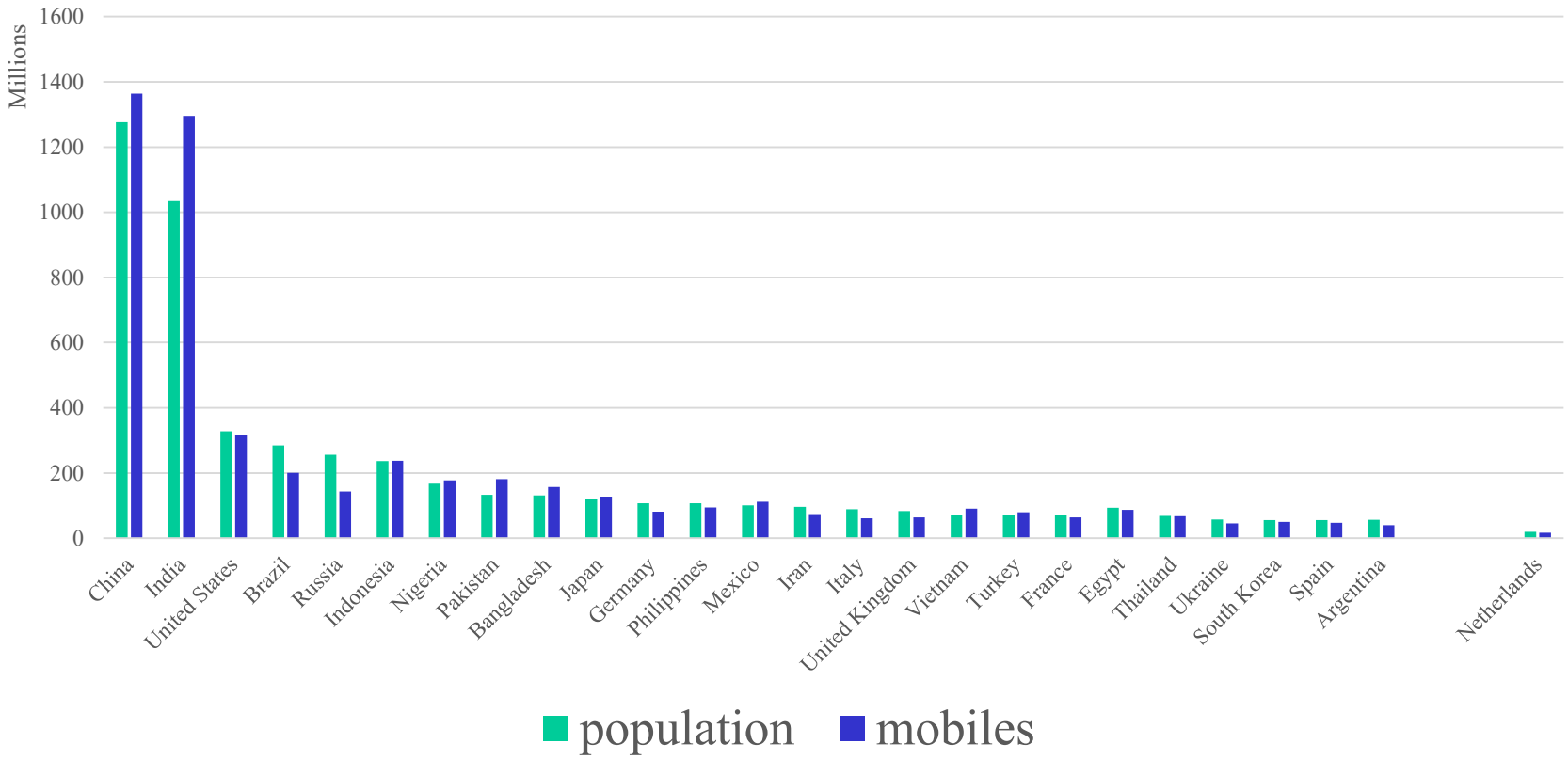


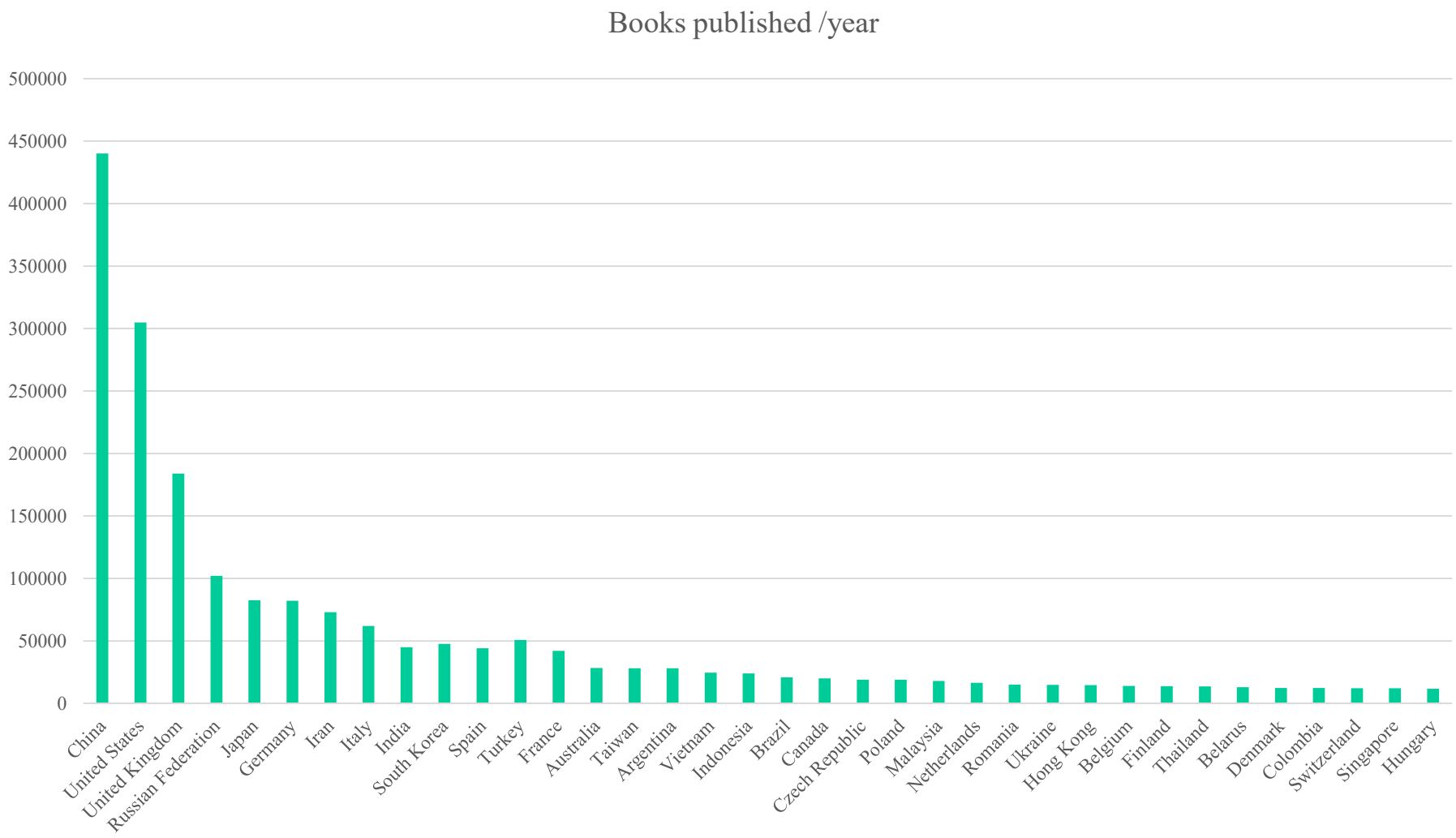
Source: Internet World Stats - www.internetworldstats.com/stats9.htm
433,651,012 estimated EU Internet users in June 2017

UNESCO Endangered Languages



Communication …Population & Mobile phones




Books published /year
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
China
United States
United Kingdom
Russian Federation
Japan
Germany
Iran
Italy
India
South Korea
Spain
Turkey
France
Australia
Taiwan
Argentina
Vietnam
Indonesia
Brazil
Canada
Czech Republic
Poland
Malaysia
Netherlands
Romania
Ukraine
Hong Kong
Belgium
Finland
Thailand
Belarus
Denmark
Colombia
Switzerland
Singapore
Hungary

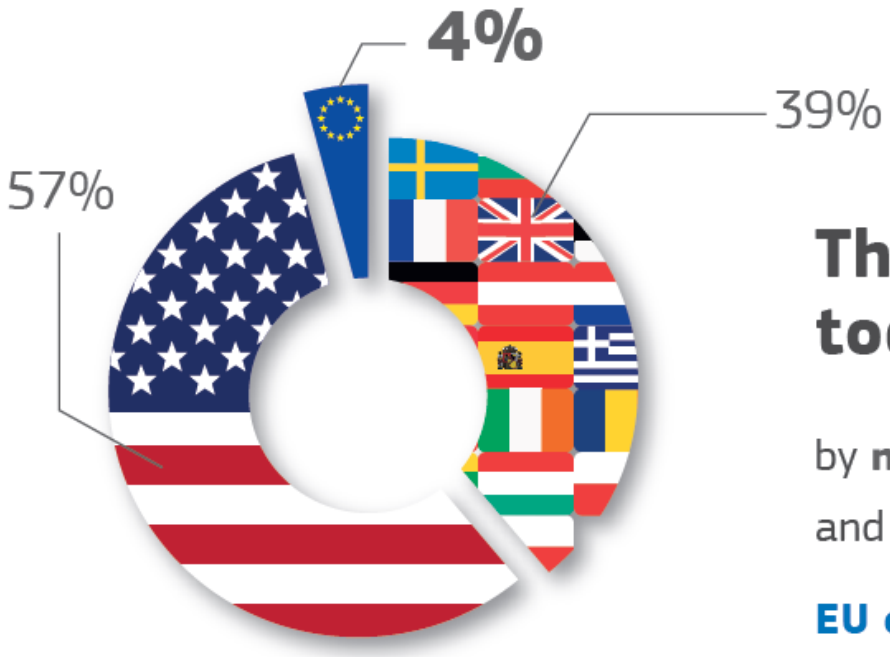
The Vision of a Digital Single Market

"Consumers need to be able
to buy the best products
at the best prices,
wherever they are in Europe."

Vice-President Ansip, Dec 2014

Accelerating growth through a connected Europe:
Speech at GSMA Mobile 360 conference in Brussels
http://europa.eu/rapid/press-release_SPEECH-14-2420_en.htm

…and the Reality



The Digital Market today is made up

by **national** online services (39%)

and **US-based** online services (57%)

EU cross-border online services represent only 4%

http://europa.eu/rapid/attachment/IP-15-4653/en/Digital_Single_Market_Factsheet_20150325.pdf

Broken already by a Simple Search

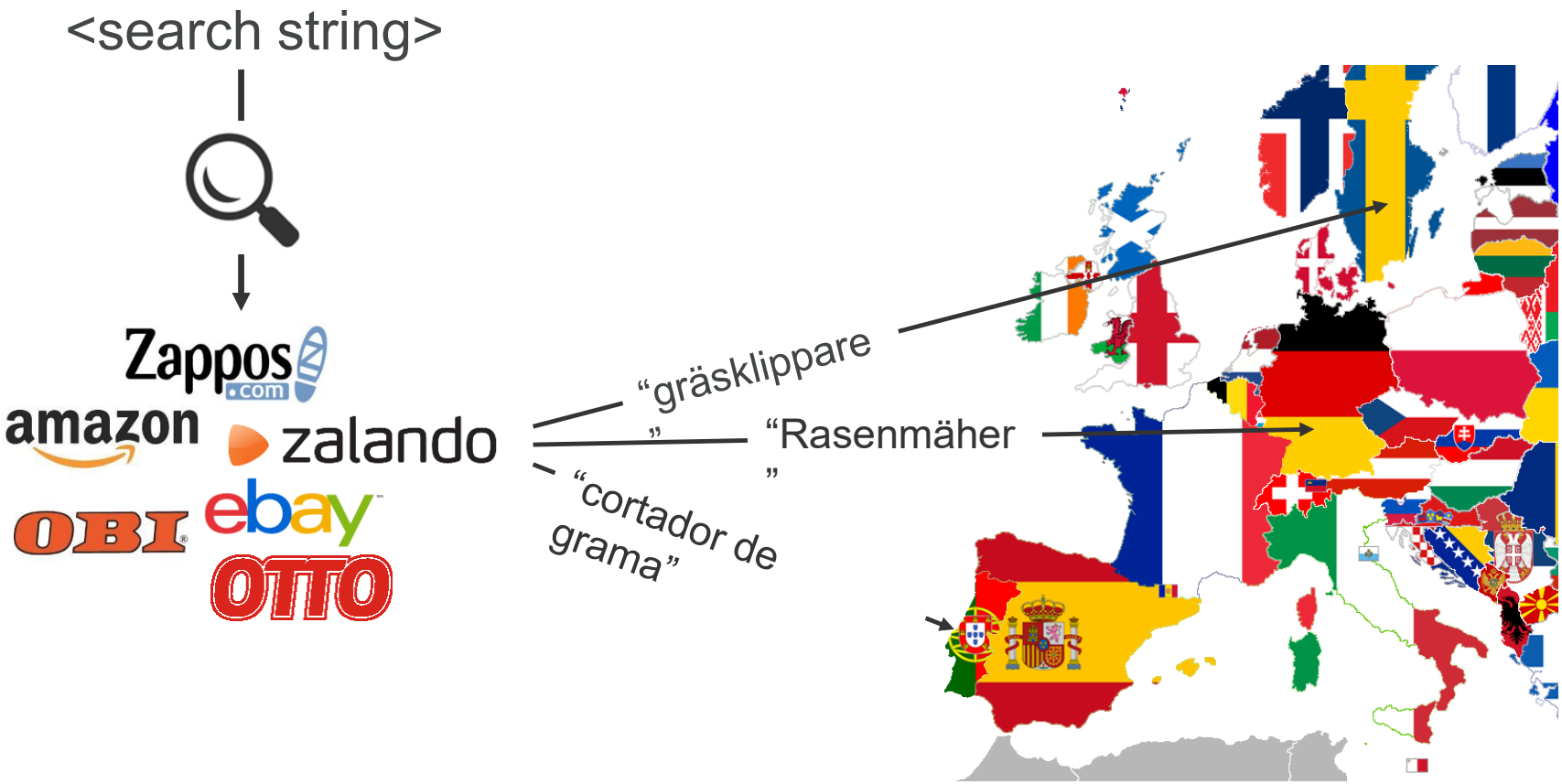
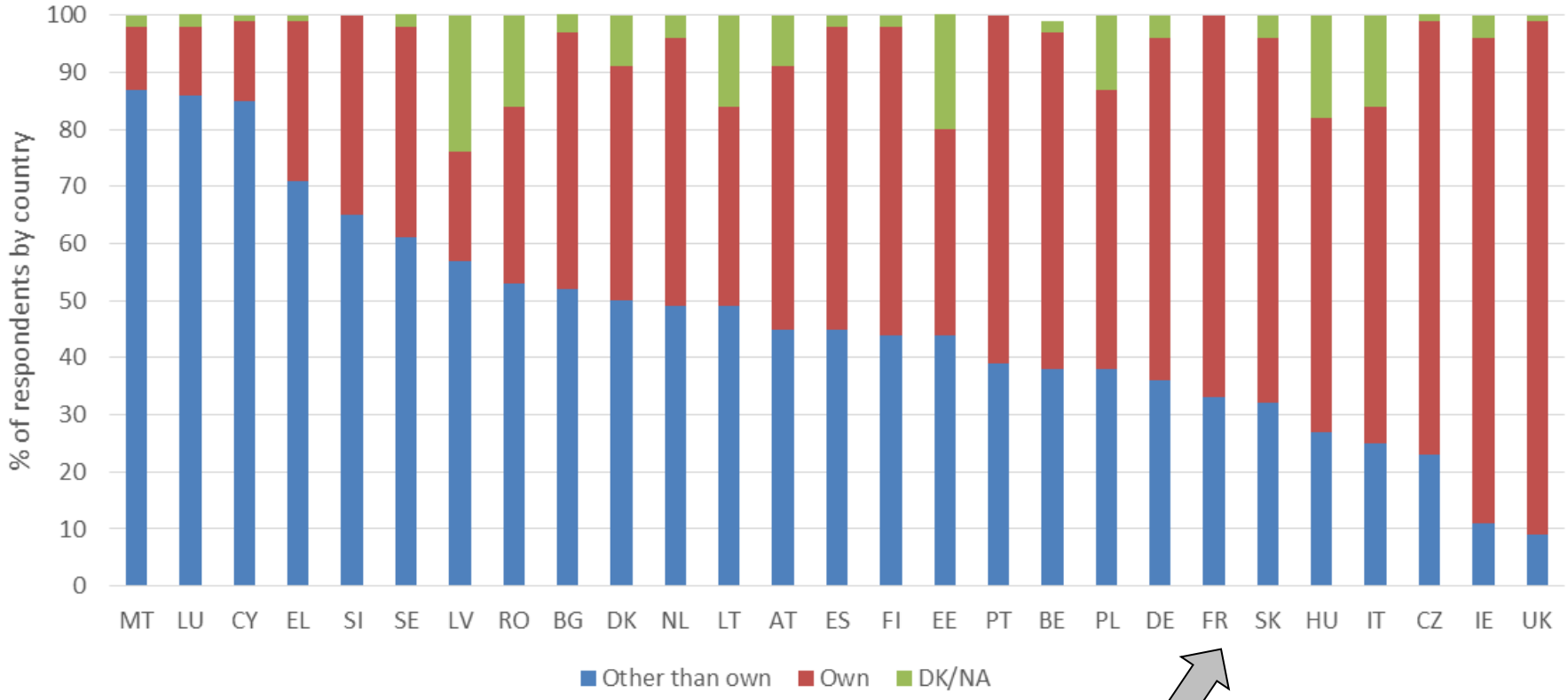
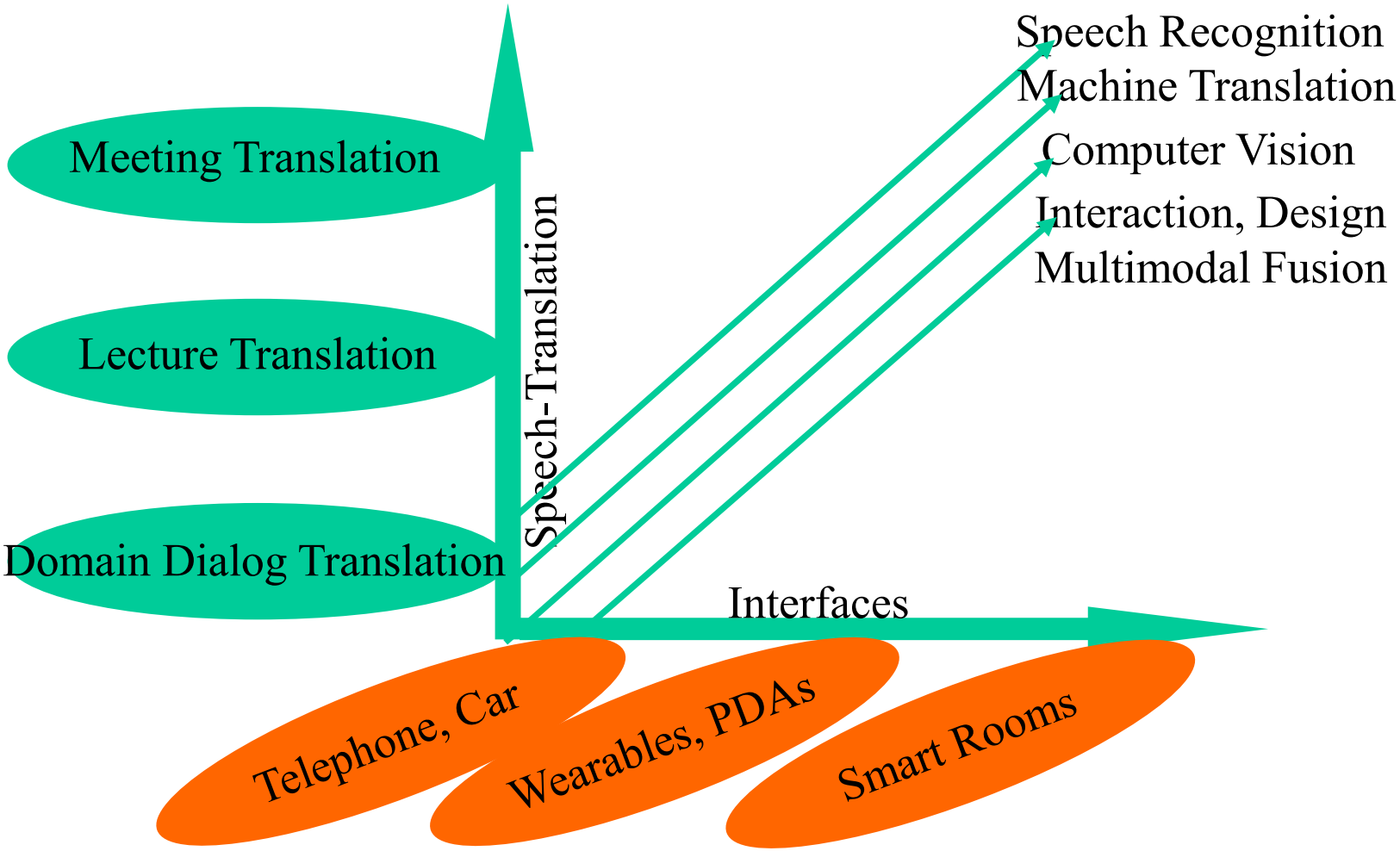


Figure 8. % of respondents by EU country using a language other than own language when writing on the Internet. Source: EC (2011: 10). Refers to writing e-mails, comments and messages posted on a website.



!42!

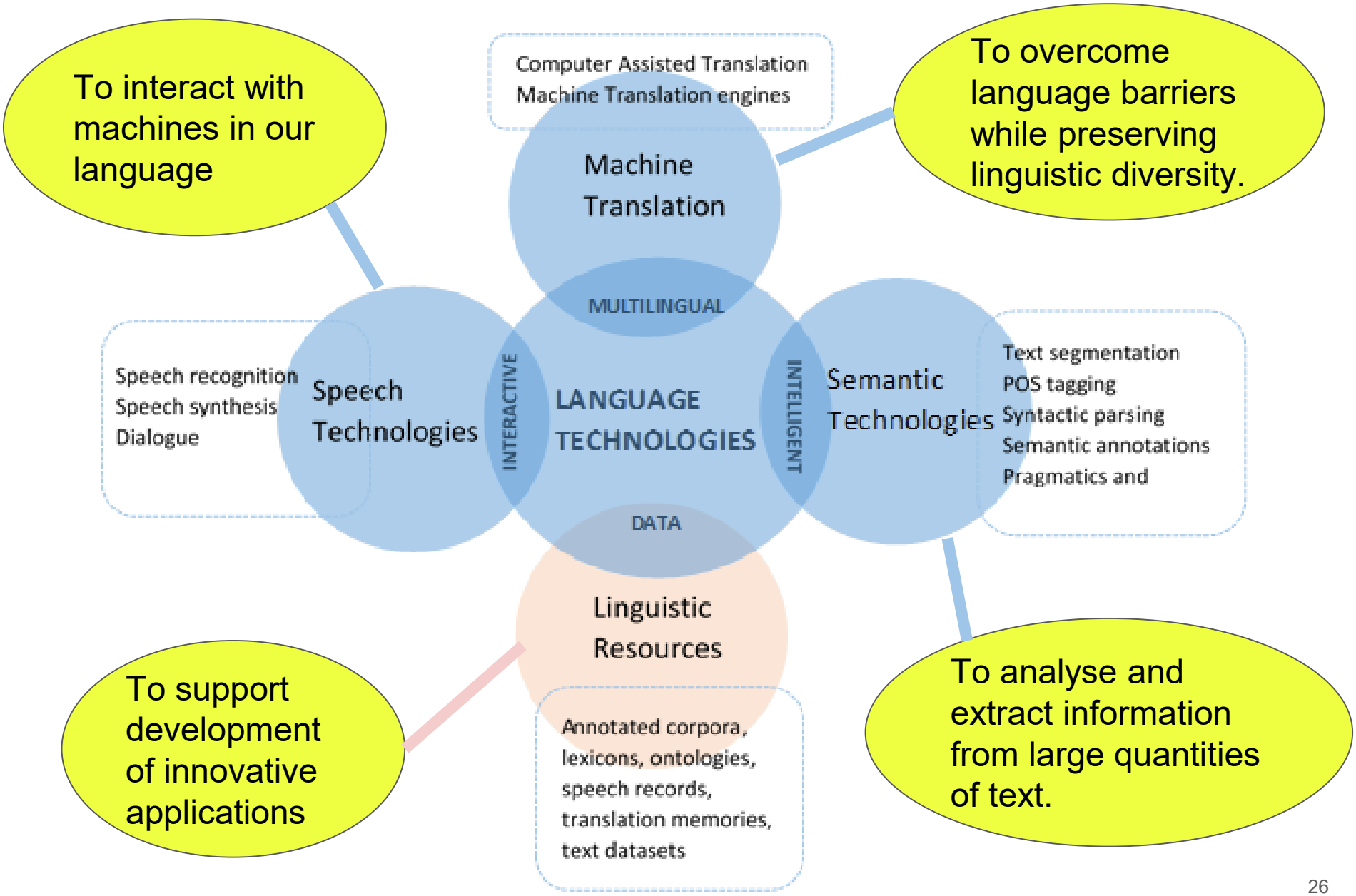
Grand Challenges // Problems



Dimensions of the Human Languages

- Speech
- Text inc. documents management (structure)
- Signs
- Hand writing
- Gestures … pointing
- Images
- Biometrics
- …. Multimodal & Multimedia
- ……

Multilinguality



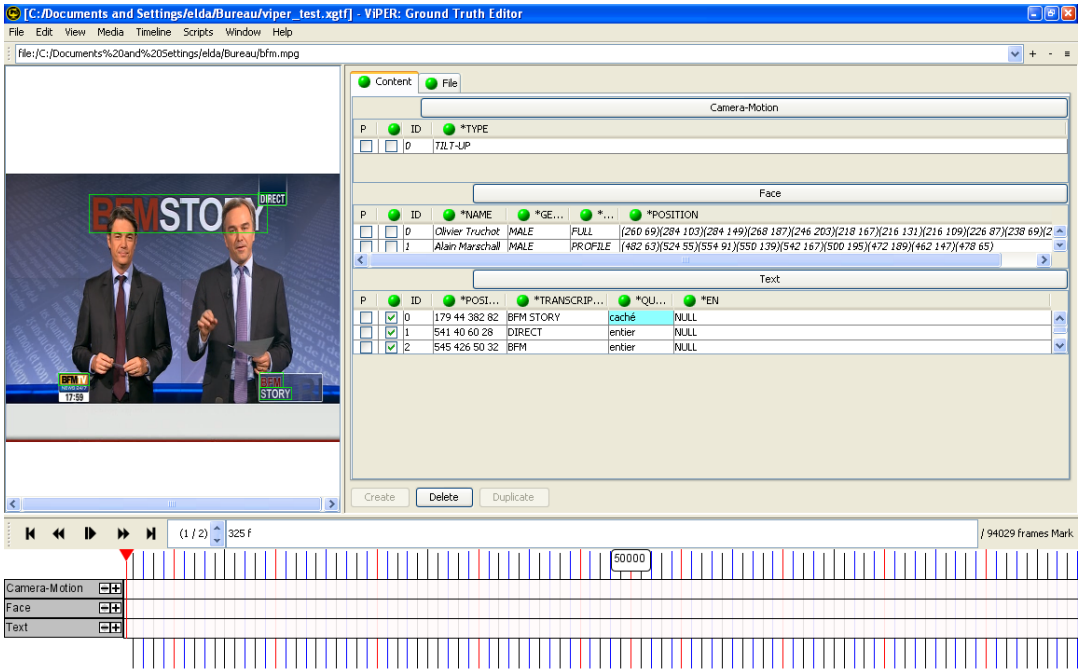
Courtesy *maite.melero@upf.edu*

Human-Machine Interactions / Mediated by Computers

- **Person localization and tracking**
- **Person identification: Face recognition, speaker identification (and fusion)**
- **Gesture recognition**
- **"attention" tracking**
- **Conversational speech recognition & understanding**
- **Acoustic scene analysis**
- **Emotion identification (facial expression, emotional features …)**
- **Topic, emotion, sentiment, ….. analytics,**
- **Speech Recognition and Understanding for dictation**
- **Speech Output (Synthesis & generation)**
- **Document classification, Text categorization**
- **… (Speech2Speech) Machine Translation**

Multimodal technologies

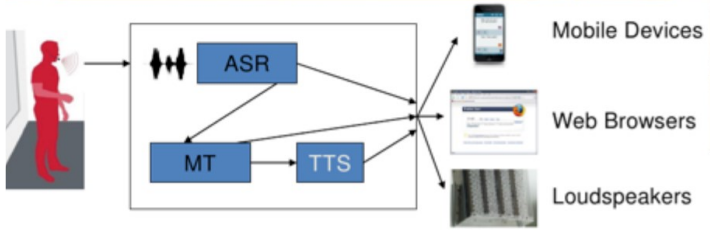
- **TV Broadcast** (REPERE project - ViPER)
  - Head localization & identification
  - Embeded text localization & transcription
  - Speech transcription & annotation



2009


JIBBIGO
Hello. Nice to meet you!
(Hello, nice to meet you.)
Hola, encantado de conocerle.
© 2009. Mobile Tech LLC

Lecture Translator in Karlsruhe



Internet Delivery

- Students bring their own Devices
- Transcription/Translation Output is Delivered via Web Page
- Interpretation Done on Server
- User Can Select Languages
- Launched 2012 as KIT Student Service
- Data Collected in Use and Evaluate

Prof. Alex Waibel

DEMO

Skype Translator

What is common to all these technologies

All are based on

MACHINE LEARNING FROM DATA

(The DATA driven Paradigm)

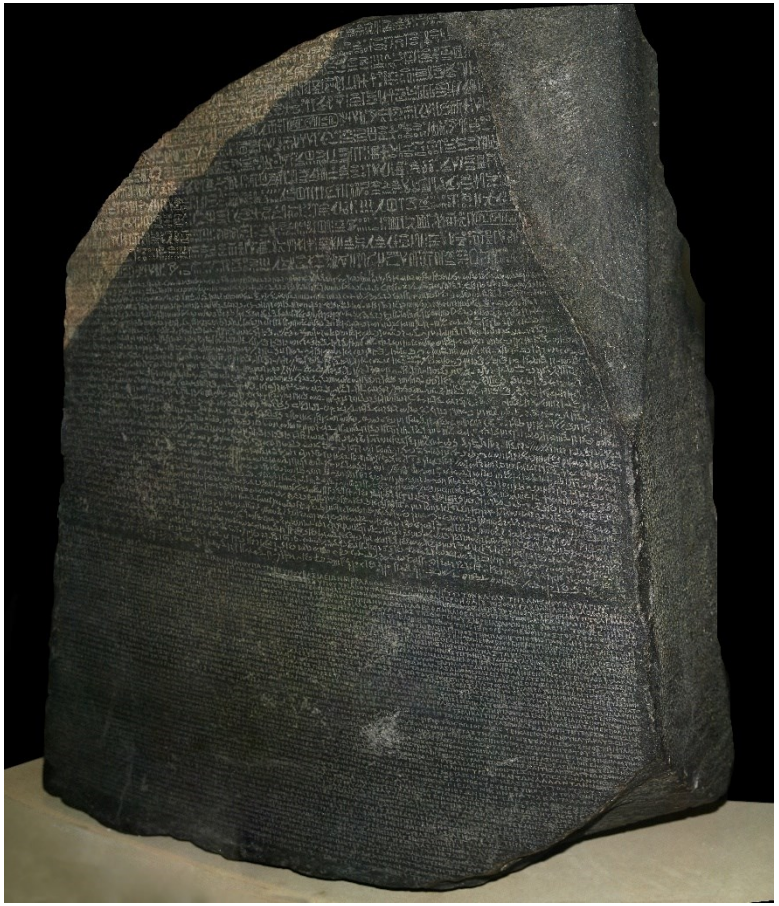
Some figures about the translation

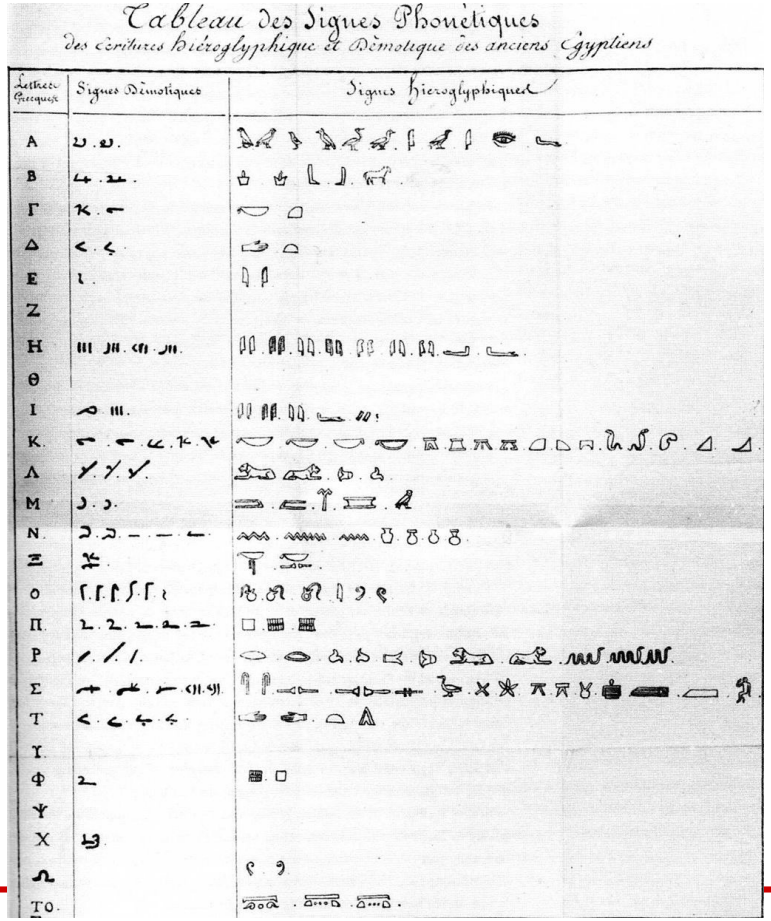
- Youtube movies and Internet video growth
  - 13h of video every ….. **MINUTE**
  - Human Transcriptions … 3-50 times (1h audio = 3h to 50h of labor)

- Translations & Interpretations
  - Over 400.000 translators (150.000 in Europe)
  - Need to translate 552 language pairs in EU, 110 in South Africa, 462 in India, (6000 languages all in all),
  - Not counting converting Sign language to/from Language A
  - Needs for translation grow by 30% …. **every year**
  - Consensus: 10% of data is translated

- Automation is essential

Champollion & The Rosetta Stone





Basis of MT

"Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.

Warren Weaver (1894-1978)

Warren Weaver and Andrew Booth

"One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'"

"Translation" (1955), in W.N. Locke and A.D. Booth (eds.),
Machine Translation of Languages (MIT Press, Cambridge, Mass.)."

Courtesy Marcelo Federico.

How does Machine Translation Work today?

Statistical MT learns from data

Two kinds of data:

- Source documents and their human translations
- Target language collections

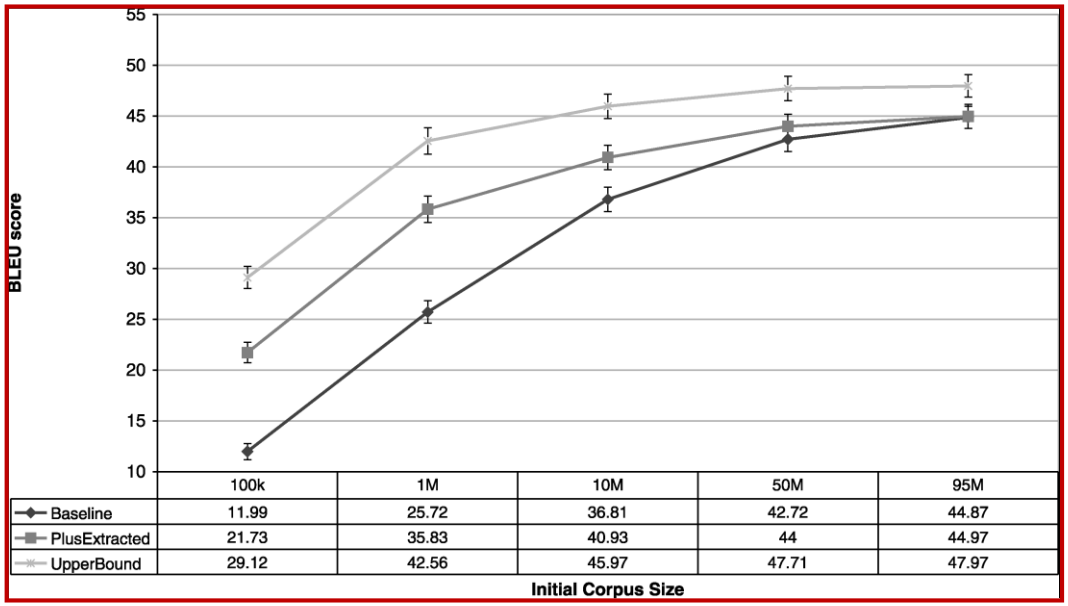
- The more data the better!
- Also: the right kind of data!

GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

Importance of data and Re-usability

✓ **Almost all technologies are data driven and based on statistical paradigms …**
(modeling based on huge amounts of date)

Let us look at MT performance when "simply" adding data



	100k	1M	10M	50M	95M
Baseline	11.99	25.72	36.81	42.72	44.87
PlusExtracted	21.73	35.83	40.93	44	44.97
UpperBound	29.12	42.56	45.97	47.71	47.97

MT performance improvements for Arabic-English
(Courtesy Dragos Stefan Munteanu and Daniel Marcu)

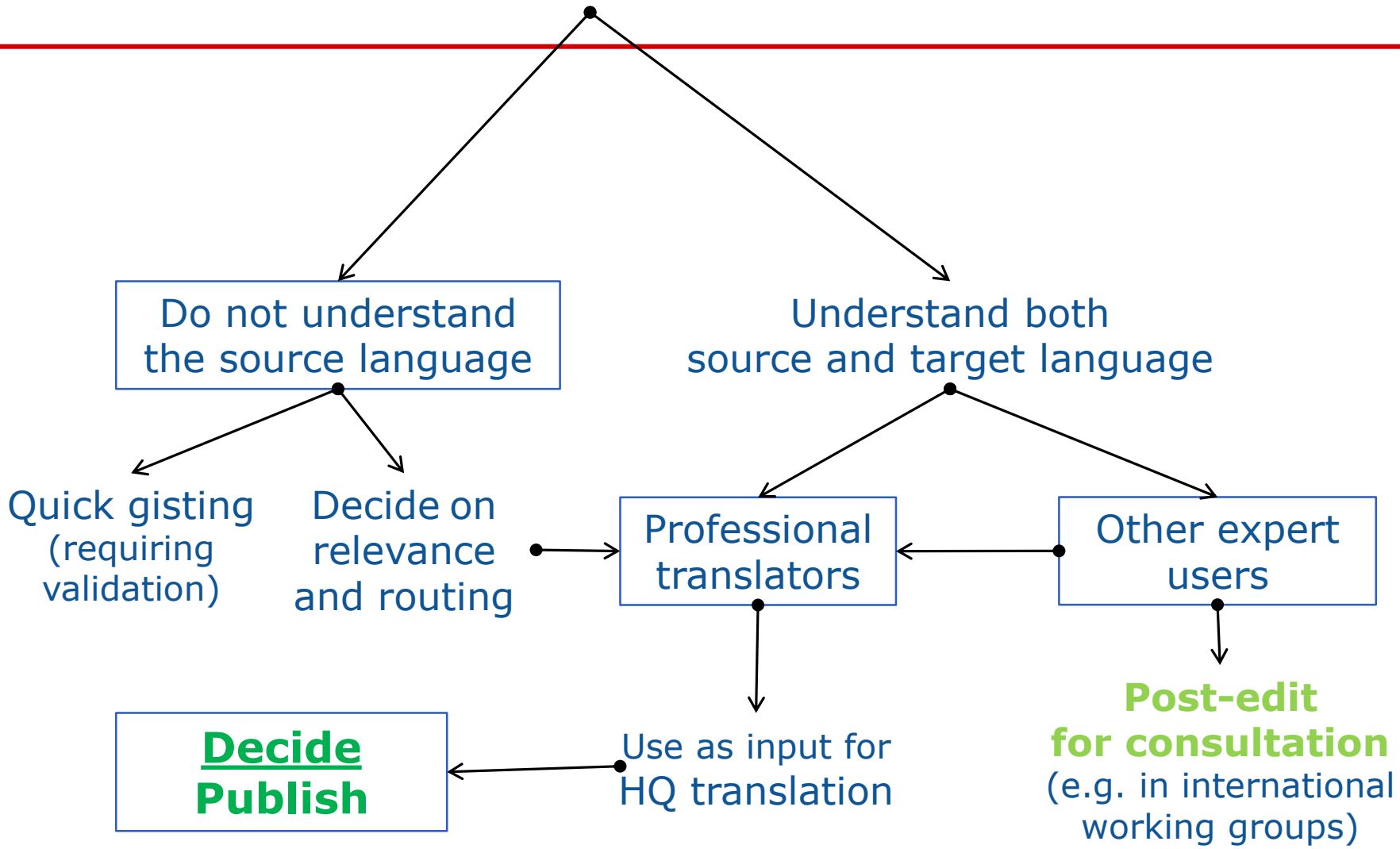
The role of Machine Translation

MT is the <u>only viable solution</u> for:

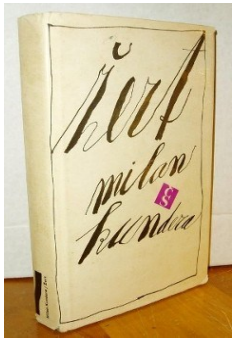
- quick and cheap access to information in foreign languages.
- understanding information received in a foreign language that otherwise could not be used or would require substantial time and costs to translate.
- making multilingual use of websites possible
- facilitating cross-lingual information search and analytics.

That is why machine translation (MT) is a critically important technology for multilingual Europe

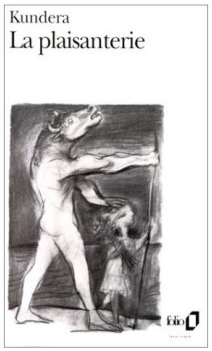
Machine Translation users



Translation is a complex process

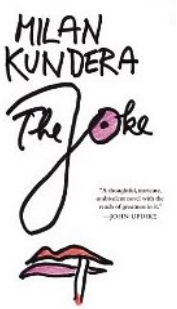


- Czech



ISBN: 978-2070366385 (1975)
- 1st French Edition

ISBN: 978-2070703739 (1985)
- 2nd French Edition



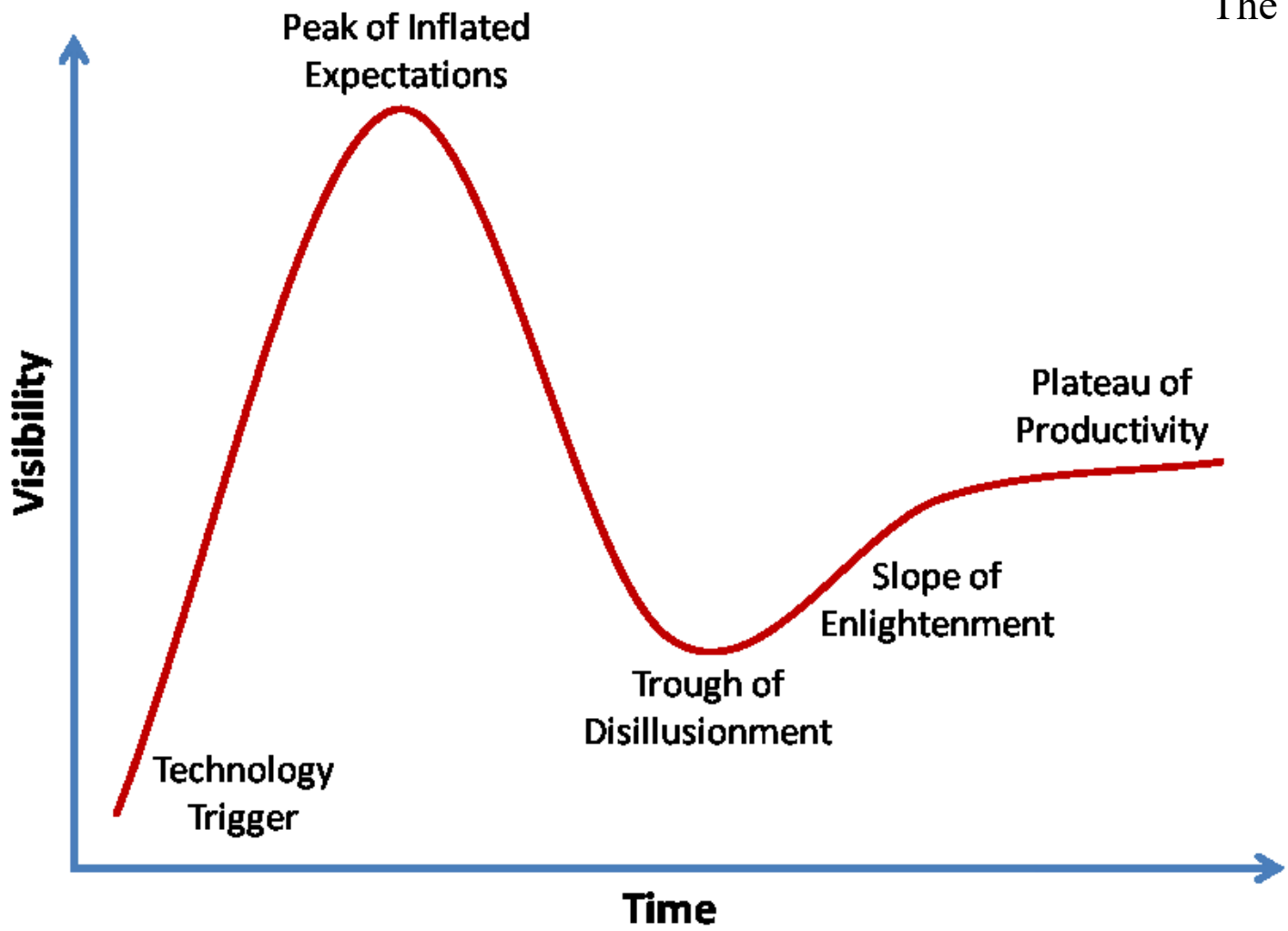
ISBN: 978-0060995058 (1993) / (1969)
- English

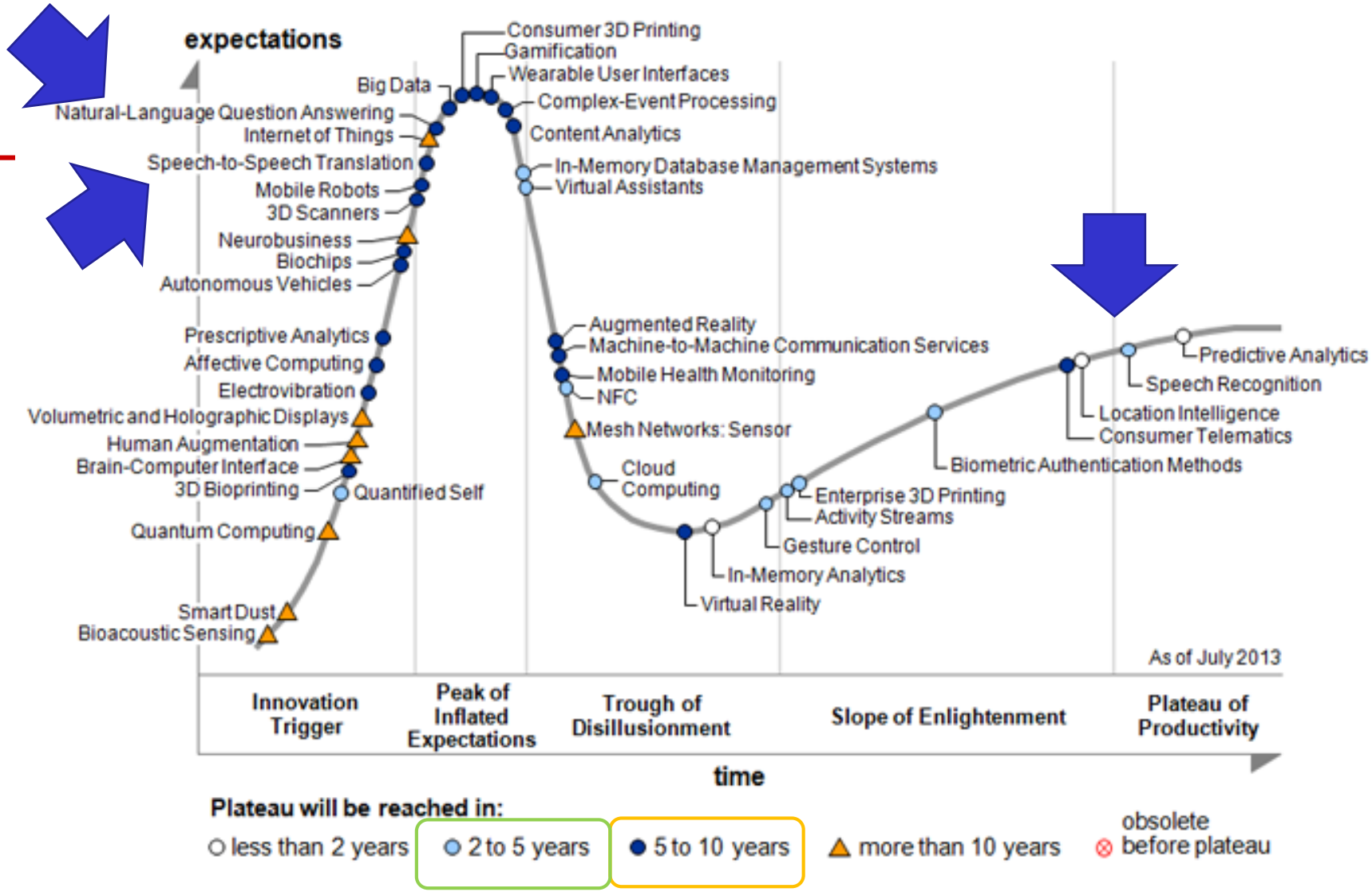
What are the trends … Challenges for the next « decade! »

1. Introduction
2. Languages: instruments of cultures, identity and business
3. Language Technologies: some examples, illustrations
4. Special focus on Automated Translation
   - Automated/Machine Translation , need for Language Resources (Data sets)
   - The MT@EC and the next generation (CEF-AT)
   - How can we help to improve it
5. What are the trends … Challenges for the next « decade! »
6. Quick conclusions
7. Q/A Session

Where do we stand today … techno trends

The Gartner Hype Cycle





Source: Gartner August 2013

The 2013 Emerging Technologies Hype Cycle highlights technologies

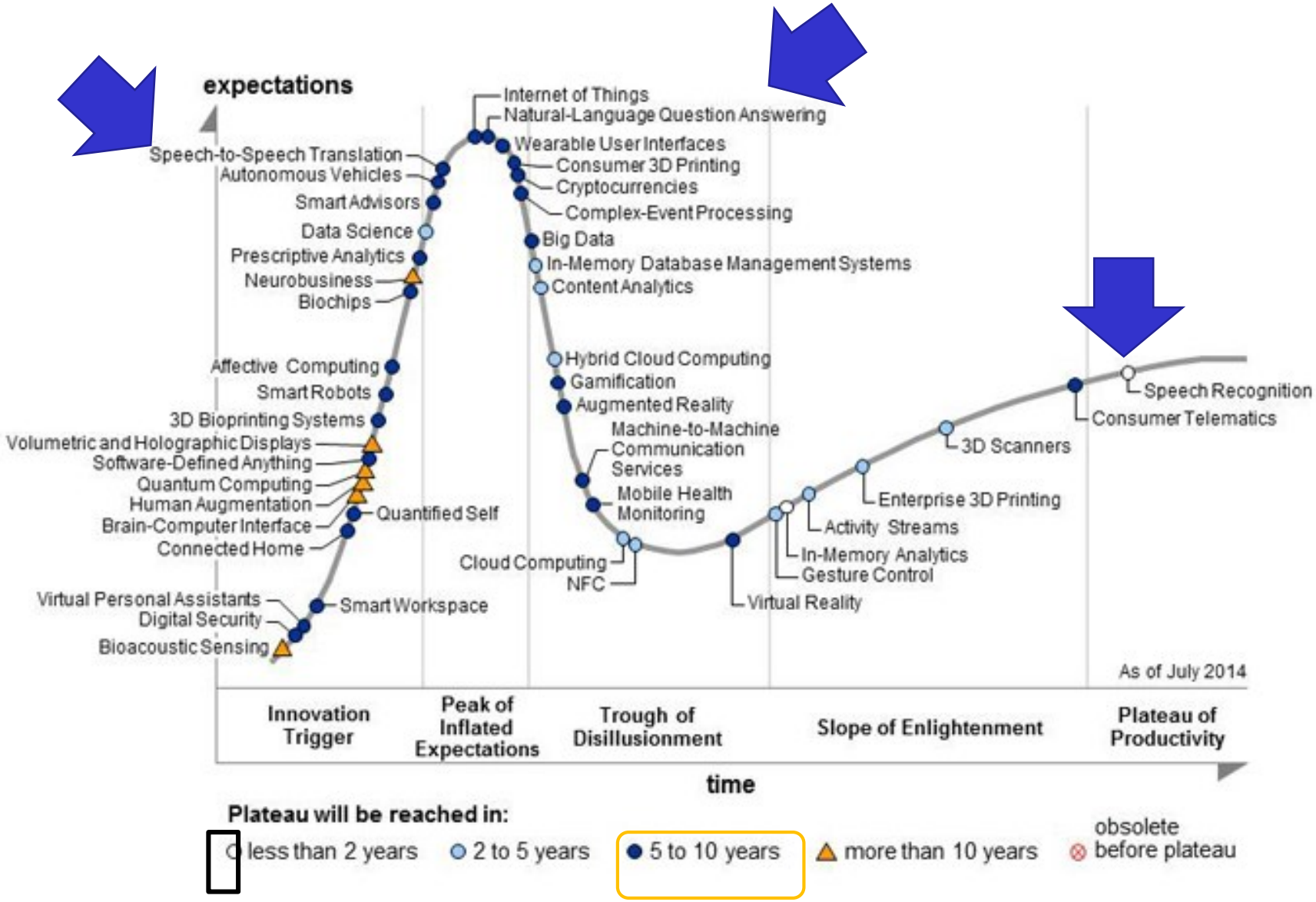
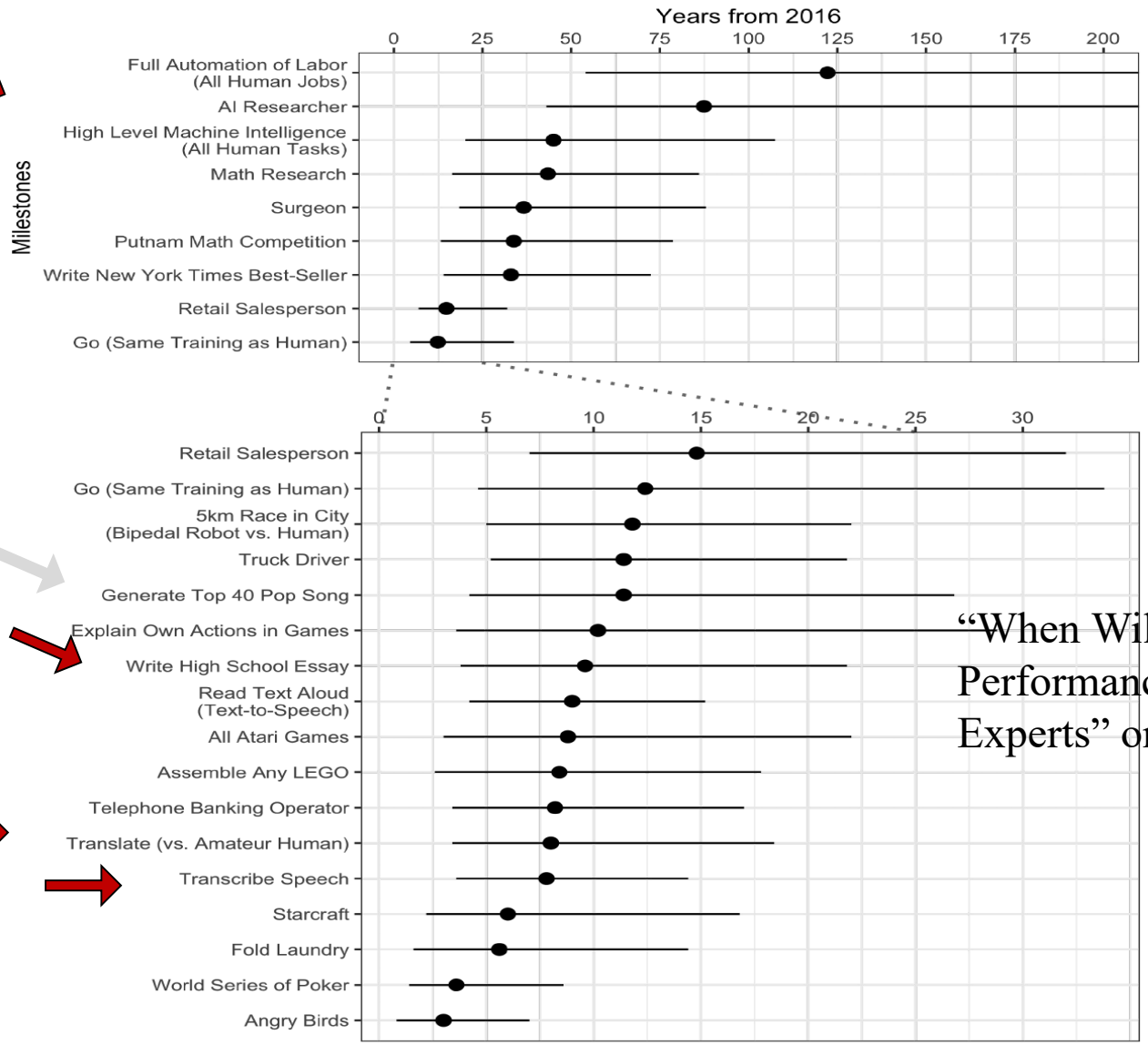


Figure 1. Hype Cycle for Emerging Technologies, 2014

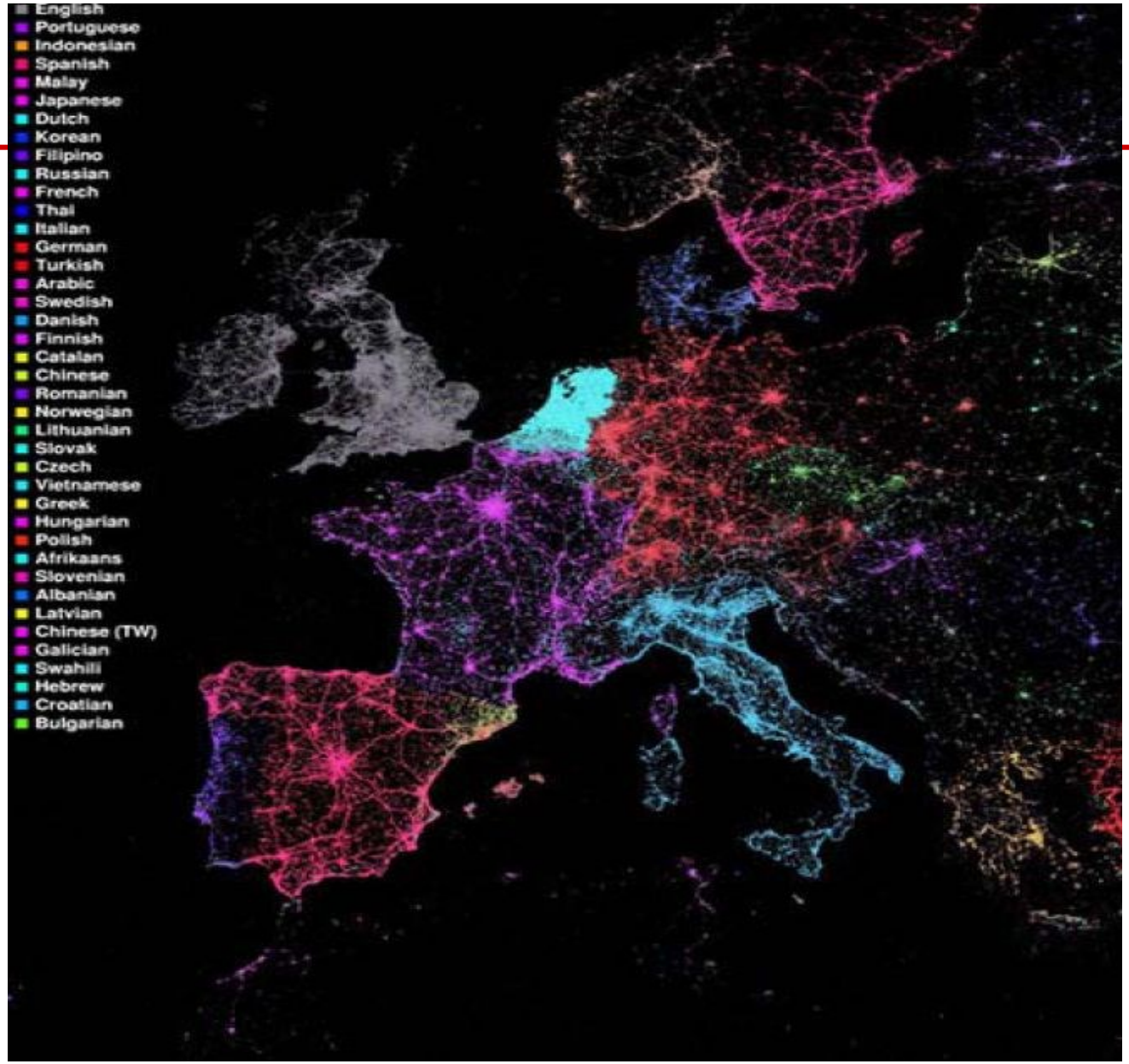


https://slator.com/academia/yale-oxford-enter-business-predicting-end-human-translator/

"When Will AI Exceed Human Performance? Evidence from AI Experts" on May 24, 2017.

Challenges … Trends ….

- More (and more) Data
- New techniques Neural Networks approaches (less data needed to start)
- Other technologies as seen in the Gartner Hype Cycle will emerge (e.g. affective computing)
- More work on the less resourced languages (those without market & lucrative segments)
  - Only 400 languages have more than 1M speakers


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

European Language Resource Coordination
Connecting Europe Facility

The European Parliament voted in favor of the resolution on "**Language equality in the digital age**" on 11. September

European Language Resource Coordination
Connecting Europe Facility

European

"18. Calls on the Commission and the Member States to develop strategies and policy action to facilitate multilingualism in the digital market; requests, in this context, that the Commission and the Member States define the minimum language resources that all European languages should possess, such as data sets, lexicons, speech records, translation memories, annotated corpora and encyclopaedic content, in order to prevent digital extinction;"

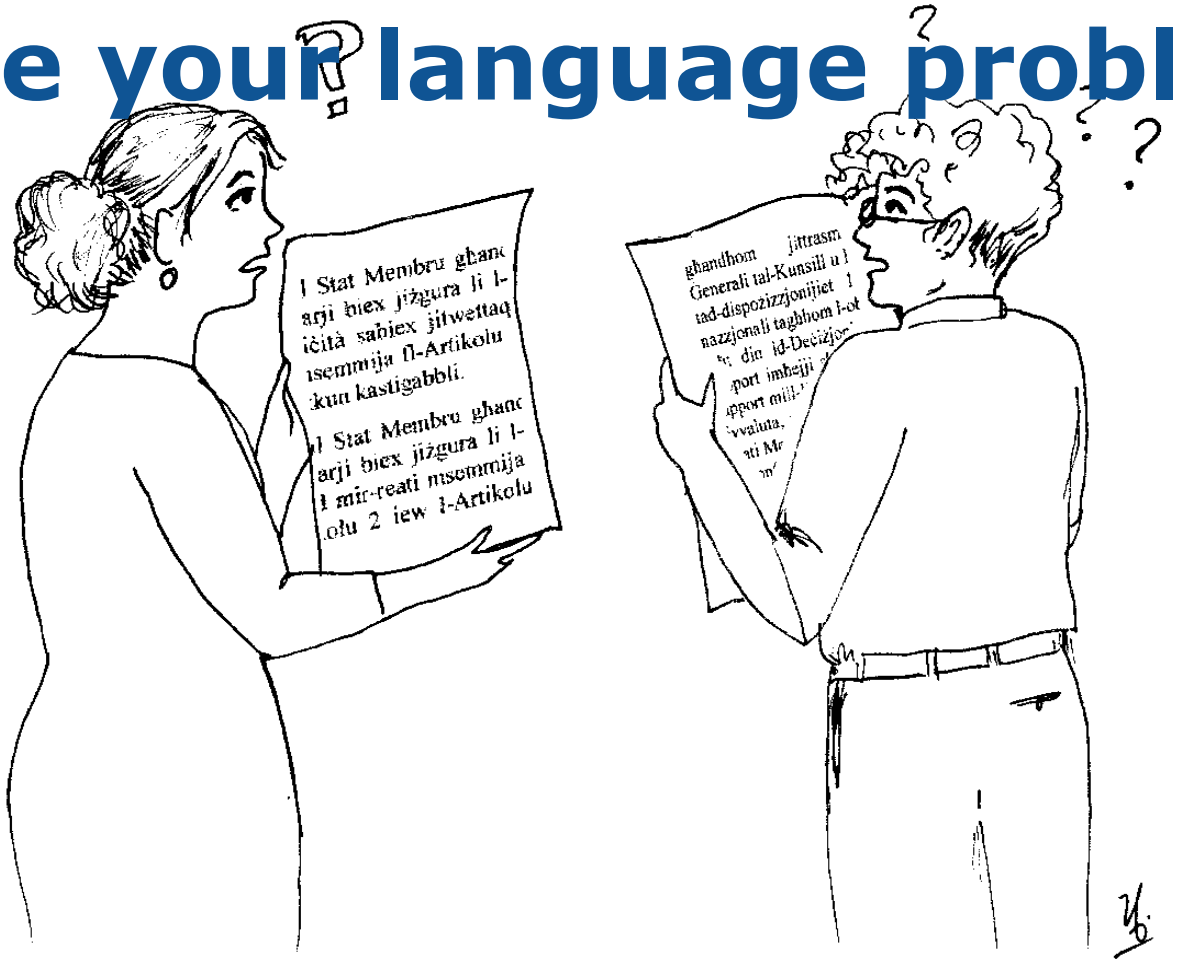
European Language Resource Coordination
Connecting Europe Facility

European

"21. Calls on the Commission to make as a priority of language technology those Member States which are small in size and have their own language, in order to pay heed to the linguistic challenges that they face;"

European Commission

Solve your language problems!



2019 | INTERNATIONAL YEAR OF
Indigenous Languages

https://en.unesco.org/news/unesco-launches-website-international-year-indigenous-languages-iyil2019

Draft prospective for a World Summit on

"Language Technologies for All (LT4All)"

in the framework of the

UNESCO Year of the Indigenous Languages – 2019

[ISCA/ELRA SIG-UL: Special Interest Group on Under-resourced Languages]

29/01/19 LT4All Meeting @ ELRA 1

Oh! One last slide ...

What happens to translators?

- Machine Translation makes mistakes (like humans) – important texts must always be checked and validated by people!
- Translation automation mainstreams multilingualism: people grow used to finding all content in their own languages.
- Translation automation multiplies translation volumes to levels never seen or dreamed of before.
- There is an increasing need for language professionals...
- ...and their job descriptions get more interesting and varied, new professions emerge:
  - High Quality translations, quality control, content management, editing, stakeholder relations, corpus linguistics, computational linguistics, configuration and optimization of translation systems...

- Translators have a bright future!


thank you
danke
謝謝
рахмат
ngiyabonga
شكرا جزيلا
teşekkür ederim
Баярлалаа
спасибо
gracias
tapadh leat
dank je
bedankt
dziękuję
obrigado
merci
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
감사합니다
ありがとう
grazie
arigatô
takk
dakujem
merci
мерси
хвала
enkosi
hvala
mauruuru
köszönöm
sagolun
taiku
ευχαριστώ
xièxie
merci

Vision

European Commission

Wouldn't it be great if I could start using a public service in any Member State from any place and obtain the information in my mother tongue?

The CEF eTranslation platform @ work

Markus Foti
eTranslation Programme Manager
DGT R.3

eTranslation platform at a glance

MT@EC

- Launched June 2013
- Legalese
- Statistical (Moses)

eTranslation

- Launched July 2017 (webservice for snippets), Nov. 2017 (web page for documents)
- Cloud based
- Neural engines

CEF.AT

- More NLPs (transliteration, named entity recognition…)
- Generic services
- Projects (ELRC, market research)

eTranslation platform at a glance

Available for:

- **individuals (submit documents through a web page)**
- **machine-to-machine use**

Users:

- **Digital Service Infrastructures (EESSI, ODR, Open Data Portal, Europeana, etc.)**
- **System suppliers (EURLex, N-Lex, Internal Market Information system…)**
- **Individuals in public administrations**

Benefits:

- **Increase speed and productivity**
- **Reduce costs**
- **Facilitate information exchange**

European Commission

MT@EC - Machine Translation

Markus FOTI
European Commission

About | Feedback | Help

Translate documents | Translate text | My translation requests | My settings | English | Logout

Select at least one source file : Choose files... **OR** Drag files here.

Translate from...

Translate into...

Domain ALL

Output format: Same as source | TMX | XLIFF

E-mail me my translation.

Delete after download.

Translate document

Selected files

Connecting Europe
eTranslation

Markus FOT
European Commissio

About | Hel

Translate documents | Translate text | My translation requests | My settings | English | Logout

Drop files to upload (or click)

Supported formats:

From *

To *

Domain SPD

Output format Same as source | TMX | XLIFF

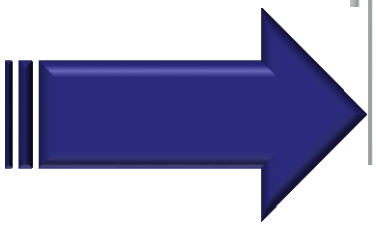
E-mail me my translation

Translate document

European Commission

One original, many translations…

English



Français

Deutsch

Italiano

Português

Polski

Lietuvių kalba

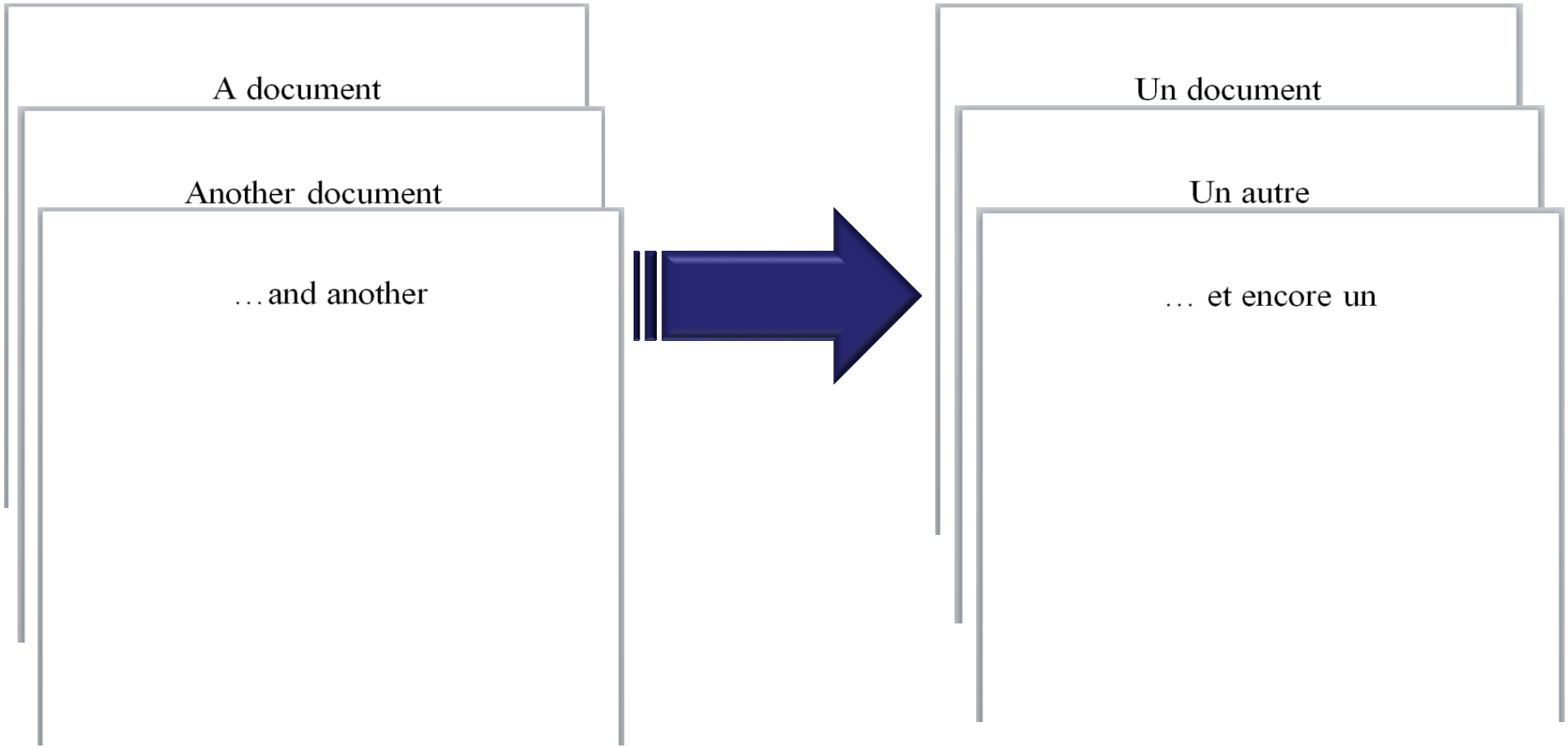
Ελληνικά

Български език

Svenska

All 24 EU languages

Many documents at once…



European Commission

eTranslation protects your privacy

- All documents deleted after 24 hours or after delivery (on demand)
- IPR rights transferred
  (Your translation belongs to YOU!)

European Commission

No one

looking

at your translations!

How to Connect to eTranslation

Open to DSIs and public administrations

- **Contact CEF-AT@ec.europa.eu with your request and use case**
- **We will provide the technical documentation on how to connect**
  - SOAP request or RESTful interface
- **Contact us for credentials**
- **Adapt your service to multilingual use!**

Behind the scenes: how it works

Statistical Machine Translation

- **MT@EC built only on "EU translations" (Euramis database)**
- **Covers all 24 languages**
- **Euramis (and MT@EC) cover EU policies, subjects and language but limited everyday language**
- **Moses-based engines**
- **Performs best when trained on large volumes of text pairs (source-translated) in specific domains**

Neural Machine Translation (NMT)

What is NMT?

- **Machine learning: artificial neural networks trained on existing translations**
- **The computer devises its own rules on how to translate**
- **Radical departure from the phrase-based SMT approaches**

Why is it important?

- **Translations read better: more fluent and grammatical**
- **Better able to fill in gaps in data used for training training**
- **Better for highly inflected languages (e.g. German, Hungarian)**

This is where the field is headed

- **Will SMT become obsolete?**

Language resources: the key to success

More data for all languages

Better lexical coverage for all languages

Machine translation adapted to your domain/subject matter

CEF.AT brings…

Reliable and trustworthy translation for EU and National Public Administrations

Support for languages with fewer speakers

Opportunities for private sector, through grants and generic services

A higher profile for language technologies, thus fostering demand

Public availability of data collected by ELRC