Technologies for Breaking Language Barriers in Europe

Georg Rehm

DFKI, Germany
georg.rehm@dfki.de

Stelios Piperidis

ILSP, R.C. “Athena”
spip@ilsp.gr

Beyond Language Barriers – ECSPM Symposium
Athens, Greece – December 01, 2017

META-NET has received funding from the EU’s Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357). Formerly co-funded by FP7 and ICT PSP through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD (grant agreement no.: 270899).

□ META-NET

60 research centres in **34** countries.
Chair of Executive Board: Jan Hajic (CUNI)
Dep.: J. van Genabith (DFKI), A. Vasiljevs (Tilde)
General Secretary: Georg Rehm (DFKI)

□ META

Multilingual Europe
Technology Alliance.
826 members in
67 countries

Iceland
Sweden
Norway
Gulf of Bothnia
Finland
Baltic Sea
Estonia
Latvia
North Sea
Denmark
Lithuania
United Kingdom
Ireland
Netherlands
Belarus
Poland
Belgium
Germany
Czech Republic
Celtic Sea
Ukraine
Moldova
Hungary
France
Bay of Biscay
Croatia
Romania
Serbia
Italy
Bulgaria
Black Sea
Spain
Portugal
Tyrrhenian Sea
Greece
Turkey
Mediterranean Sea
Tunisia
Morocco
Israel

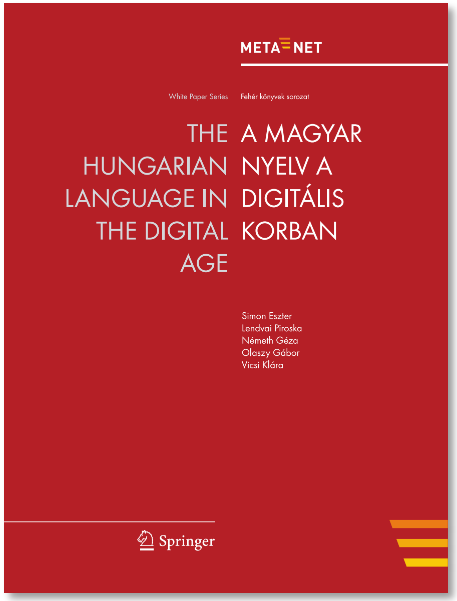
(published in 2013)

META-NET

STRATEGIC
RESEARCH
AGENDA FOR
MULTILINGUAL
EUROPE 2020

edited by the
META Technology Council

(31 volumes; published in 2012)


META-NET

White Paper Series

Fehér könyvek sorozat

THE
HUNGARIAN
LANGUAGE IN
THE DIGITAL
AGE

A MAGYAR
NYELV A
DIGITÁLIS
KORBAN

Simon Eszter
Lendvai Piroska
Németh Géza
Olaszy Gábor
Vicsi Klára

Springer

T4ME (META-NET) CESAR META-NORD METANET4U

- **META-FORUM 2017** – November 13/14, Brussels, Belgium

  Towards a Human Language Project
- **META-FORUM 2016** – July 04/05, Lisbon, Portugal

  Beyond Multilingual Europe
- **META-FORUM 2015** – April 27, Riga, Latvia

  Technologies for the Multilingual Digital Single Market
- **META-FORUM 2013** – September 19/20, Berlin, Germany

  Connecting Europe for New Horizons
- **META-FORUM 2012** – June 20/21, Brussels, Belgium

  A Strategy for Multilingual Europe
- **META-FORUM 2011** – June 27/28, Budapest, Hungary

  Solutions for Multilingual Europe
- **META-FORUM 2010** – November 17/18, Brussels, Belgium

  Challenges for Multilingual Europe

CRACKER

Cracking the Language Barrier

Coordination, Evaluation and Resources for European MT Research

Coordination and Support Action, H2020-ICT17, 2015–2017, 36 months – http://www.cracker-project.eu

1	DFKI	Germany	Georg Rehm
2	CUNI	Czech Republic	Jan Hajic
3	ELDA	France	Khalid Choukri
4	FBK	Italy	Marcello Federico
5	ATHENA RC	Greece	Stelios Piperidis
6	UEDIN	UK	Philipp Koehn
7	USFD	UK	Lucia Specia

DFKI

UNIVERZITA KARLOVA V PRAZE

eL DA

FONDAZIONE BRUNO KESSLER

THE UNIVERSITY of EDINBURGH

Ερευνητικό Κέντρο Αθηνά

Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες της Πληροφορίας, των Επικοινωνιών, της Γνώσης

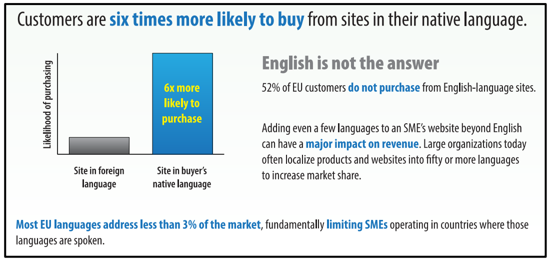
The University Of Sheffield.

Communities

- META-NET incl. META-SHARE and META
- MT evaluation initiatives – WMT, IWSLT, MT Marathons
- MT and other LT industry
- Language resources – META-SHARE, ELRA
- HT/MT evaluation tools – translate5
- Translation industry and translation profession
- MT user communities

Strategic Agenda for the Multilingual Digital Single Market

- Version 0.5 presented at Riga Summit 2015.
- Version 0.9 presented at META-FORUM 2016.
- Version 1.0 presented at META-FORUM 2017.



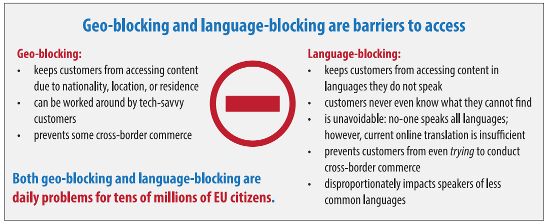
Customers are **six times more likely to buy** from sites in their native language.

English is not the answer

52% of EU customers **do not purchase** from English-language sites.

Adding even a few languages to an SME's website beyond English can have a **major impact on revenue**. Large organizations today often localize products and websites into fifty or more languages to increase market share.

Most EU languages address less than 3% of the market, fundamentally **limiting SMEs** operating in countries where those languages are spoken.



Geo-blocking and language-blocking are barriers to access

Geo-blocking:

- keeps customers from accessing content due to nationality, location, or residence
- can be worked around by tech-savvy customers
- prevents some cross-border commerce

Language-blocking:

- keeps customers from accessing content in languages they do not speak
- customers never even know what they cannot find
- is unavoidable: no-one speaks all languages; however, current online translation is insufficient
- prevents customers from even *trying* to conduct cross-border commerce
- disproportionately impacts speakers of less common languages

Both geo-blocking and language-blocking are daily problems for tens of millions of EU citizens.

Strategic Agenda for the Multilingual Digital Single Market

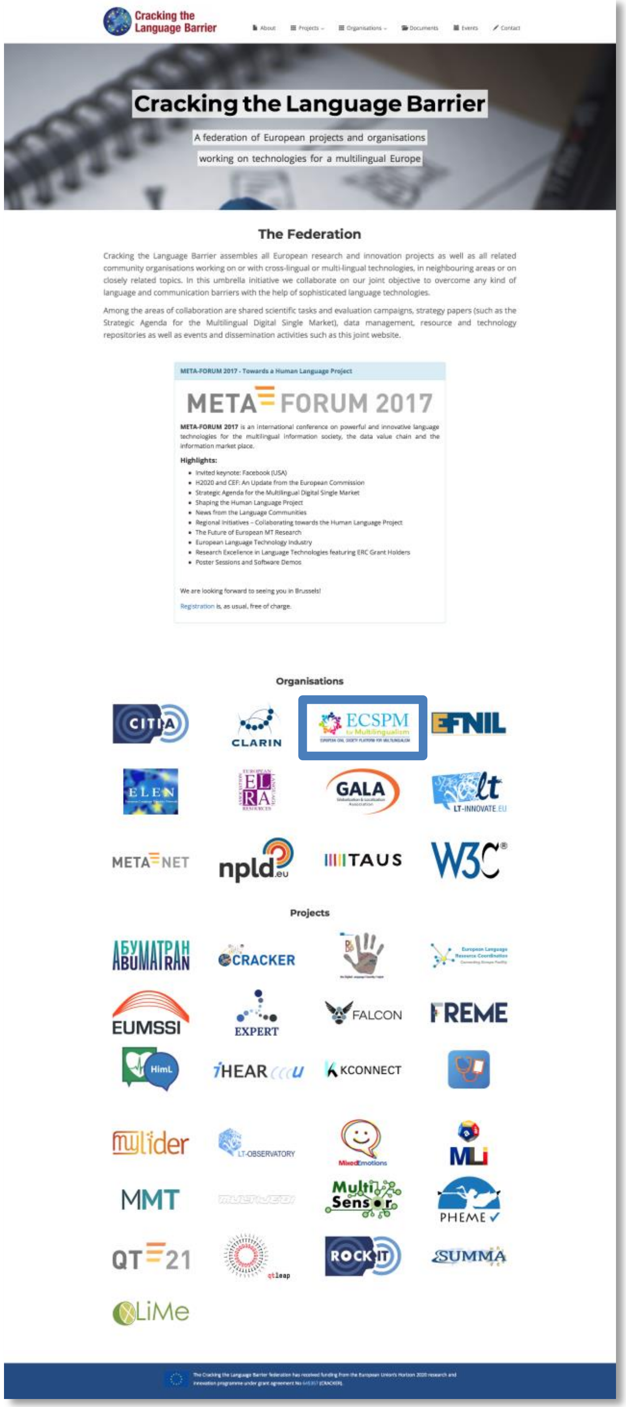
Technologies for Overcoming Language Barriers towards a truly integrated European Online Market

DRAFT

Version 0.5 – April 22, 2015

Cracking the Language Barrier

- A federation of European projects and organisations working on technologies for a multilingual Europe.
- Multi-lateral Memorandum of Understanding; 12 organisations and 25 projects on board already (including FP7 and H2020-ICT15).
- Selected areas of collaboration: data management and repositories (incl. Data Management Plan), tools, shared tasks, evaluations, events.
- Goal: provide *one umbrella* for the whole European LT community.



Cracking the Language Barrier

A federation of European projects and organisations working on technologies for a multilingual Europe

The Federation

Cracking the Language Barrier assembles all European research and innovation projects as well as all related community organisations working on or with cross-lingual or multi-lingual technologies, in neighbouring areas or on closely related topics. In this umbrella initiative we collaborate on our joint objective to overcome any kind of language and communication barriers with the help of sophisticated language technologies.

Among the areas of collaboration are shared scientific tasks and evaluation campaigns, strategy papers (such as the Strategic Agenda for the Multilingual Digital Single Market), data management, resource and technology repositories as well as events and dissemination activities such as this joint website.

META-FORUM 2017 - Towards a Human Language Project

META-FORUM 2017

META-FORUM 2017 is an international conference on powerful and innovative language technologies for the multilingual information society, the data value chain and the information market place.

Highlights:

- Invited keynote: Facebook (USA)
- H2020 and CEF: An Update from the European Commission
- Strategic Agenda for the Multilingual Digital Single Market
- Shaping the Human Language Project
- News from the Language Communities
- Regional Initiatives – Collaborating towards the Human Language Project
- The Future of European MT Research
- European Language Technology Industry
- Research Excellence in Language Technologies featuring ERC Grant Holders
- Poster Sessions and Software Demos

We are looking forward to seeing you in Brussels!

Registration is, as usual, free of charge.

Organisations

Projects

The Cracking the Language Barrier federation has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645357 (CRACKER).

http://www.cracker-project.eu • http://www.meta-net.eu

META-NET White Papers: Continued Interest

□ Basque
□ Bulgarian*
□ Catalan
□ Croatian*
□ Czech*
□ Danish*
□ Dutch*
□ English*
□ Estonian*
□ Finnish*
□ French*
□ Galician
□ German*
□ Greek*
□ Hungarian*
□ Icelandic
□ Irish*
□ Italian*
□ Latvian*
□ Lithuanian*
□ Maltese*
□ Norwegian
□ Polish*
□ Portuguese*
□ Romanian*
□ Serbian
□ Slovak*
□ Slovene*
□ Spanish*
□ Swedish*
□ Welsh

http://www.meta-net.eu/whitepapers

* Official EU language

MT

excellent	good	moderate	fragmentary	weak or no support through LT
	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, Welsh

Text Analytics

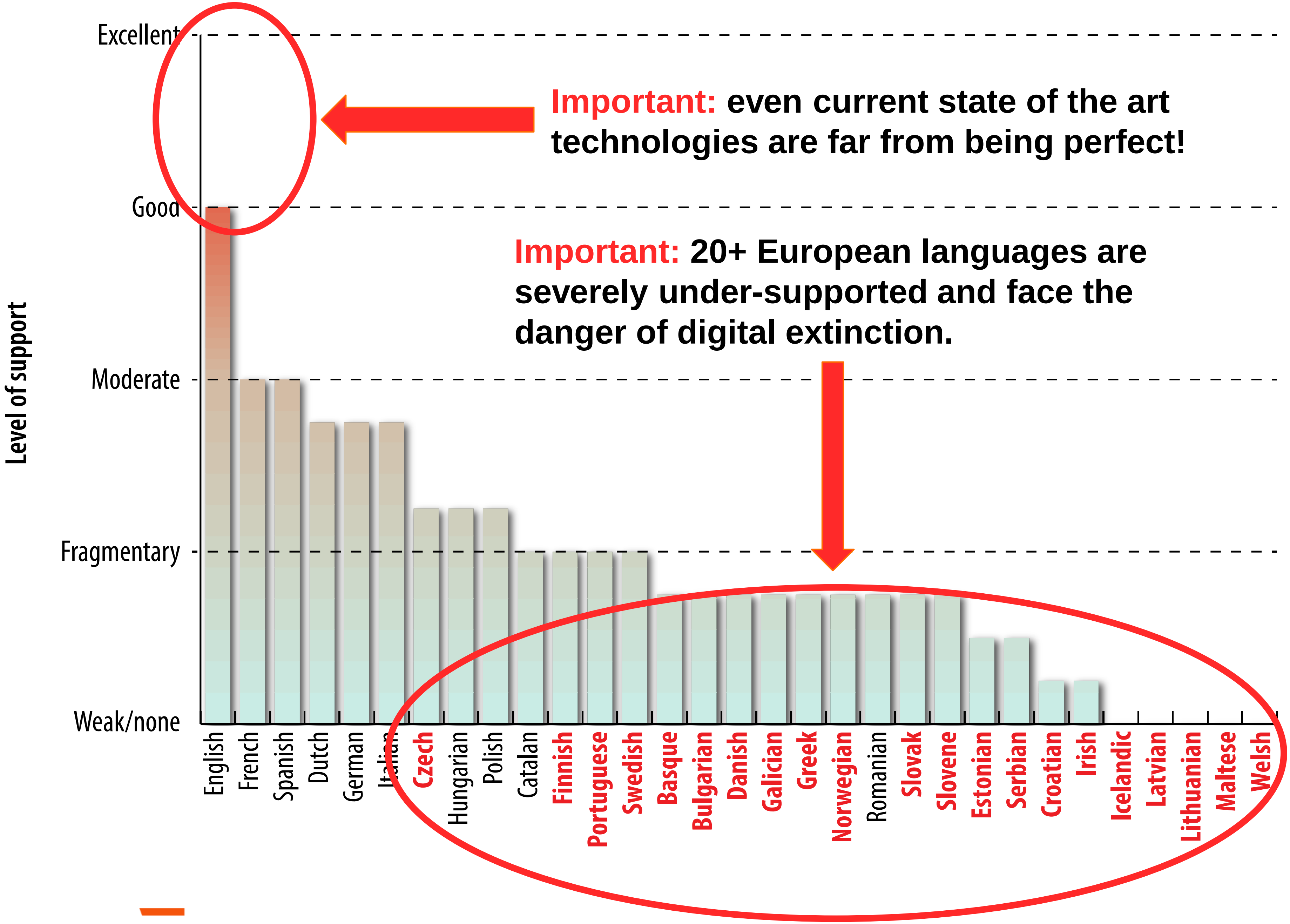
excellent	good	moderate	fragmentary	weak or no support through LT
	English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian, Welsh

Speech

excellent	good	moderate	fragmentary	weak or no support through LT
	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, Welsh

Resources

excellent	good	moderate	fragmentary	weak or no support through LT
	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, Welsh



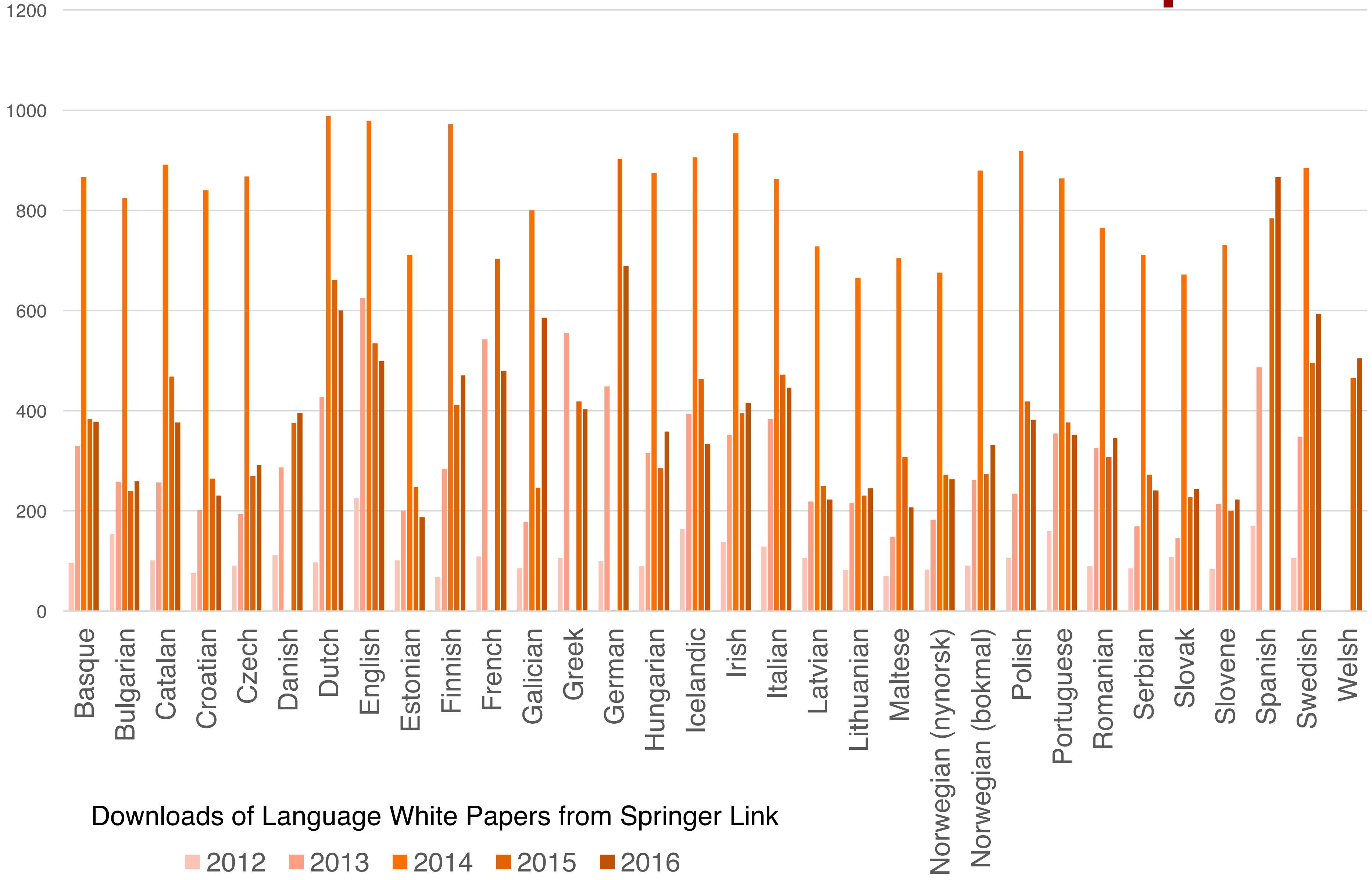
Important: even current state of the art technologies are far from being perfect!

Important: 20+ European languages are severely under-supported and face the danger of digital extinction.

META-NET

Source: META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, September 2012. Georg Rehm and Hans Uszkoreit (series editors)

Continued Interest in the White Papers



Continued Interest in the White Papers

	Basque	Bulgarian	Catalan	Croatian	Czech	Danish	Dutch	English	Estonian	Finnish	French
2012	96	153	101	76	91	111	97	225	101	68	109
2013	330	258	256	202	194	286	427	625	201	284	542
2014	866	825	891	840	868	1,063	988	978	711	972	1032
2015	383	240	468	264	269	375	661	534	248	412	703
2016	**378**	**259**	**377**	**231**	**292**	**395**	**600**	**499**	**187**	**470**	**480**

	Galician	Greek	German	Hungarian	Icelandic	Irish	Italian	Latvian	Lithuanian	Maltese	Norwegian (nynorsk)
2012	86	107	100	90	164	138	128	107	81	70	83
2013	178	556	449	315	393	352	383	218	216	148	182
2014	799	1133	1382	874	905	954	862	728	665	704	676
2015	246	419	903	285	463	395	472	250	230	307	272
2016	**585**	**403**	**689**	**358**	**334**	**416**	**446**	**222**	**245**	**207**	**263**

	Norwegian (bokmal)	Polish	Portug.	Romanian	Serbian	Slovak	Slovene	Spanish	Swedish	Welsh
2012	91	107	160	90	86	108	84	170	107	–
2013	262	234	355	326	169	145	214	486	348	–
2014	879	918	863	765	711	672	731	1195	885	1014
2015	273	419	377	307	272	228	200	784	496	466
2016	**331**	**382**	**352**	**345**	**241**	**243**	**222**	**866**	**593**	**505**

Downloads of Language White Papers from Springer Link

Multilingual Europe and the Digital Single Market

- Multilingualism is at the very heart of the European idea
- 24 EU languages – all languages have the same status
- Dozens of regional and minority languages as well as languages of immigrants and trade partners
- Economic challenges:
  - If the DSM is not multilingual, there will be 20+ isolated markets
  - Language barriers are market barriers
- Social and public challenges:
  - Empower all citizens to use their mother tongues online/offline
  - Enable cross-border, cross-lingual, cross-cultural communication
  - Provide multilingual digital public services
  - Restore trust in media (fake news debate, filter bubble issue etc.)

Strategic Agenda for Multilingual Europe

META NET

STRATEGIC RESEARCH AGENDA FOR MULTILINGUAL EUROPE 2020

edited by the
META Technology Council

- Published in early 2013.
- First strategic research agenda for our field.
- Complex process of collecting and shaping technology visions.
- Hundreds of researchers participated.
- Broad topics around Multilingual Europe in general.

META NET

History

META-NET

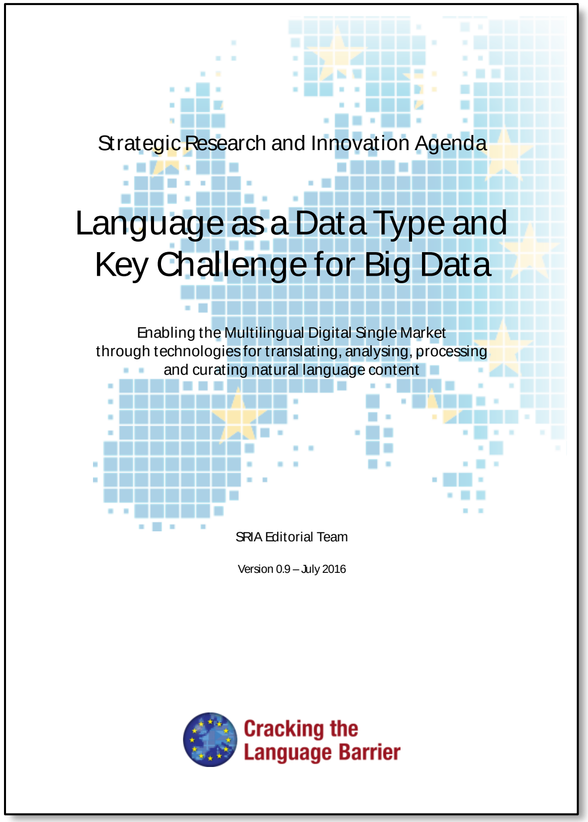
- SRIA V0.9 unveiled at META-FORUM 2016
- Prepared, presented and endorsed by the Cracking the Language Barrier federation (editorial team).
- Explains how the LT community is going to make the DSM multilingual.

Strategic Agenda for the Multilingual Digital Single Market

Technologies for Overcoming Language Barriers towards a truly integrated European Online Market

DRAFT

Version 0.5 – April 22, 2015



Strategic Research and Innovation Agenda

Language as a Data Type and Key Challenge for Big Data

Enabling the Multilingual Digital Single Market through technologies for translating, analysing, processing and curating natural language content

SRIA Editorial Team

Version 0.9 – July 2016

Cracking the Language Barrier

Cracking the Language Barrier

Application Areas in V0.9

- **Multilingual E-commerce**
  - Customer-facing vs. back-office facing (after-market, after-sales)
  - Crosslingual search, CRM, helpdesks, processes, workflows
  - Semantic, crosslingual product descriptions and catalogues
  - Online dispute resolution
- **Multilingual Content, Media, Verticals**
  - Content analytics, curation, generation (incl. authoring support)
  - Multimodal communication (speech, written, IoT)
  - Vertical domains: health, government, mobility, energy, legal.
- **Translation, Language, Knowledge, Data**
  - Translation Centre – written/spoken, automatic/human
  - Crosslingual public and social intelligence, business intelligence
  - HQ resources, under-resourced languages, domain-specific LRs

Current Developments

- **Multilingual Europe:** our languages enjoy equal status yet digital extinction of the majority of EU languages is a very severe danger.
- **Language Technology Research and Innovation in Europe:** World class research results (e.g., in QT21), strong SME base, thousands of LSPs; fragmentation; need for coordination.
- **Digitisation of our Continent – Big need for HQ Language Technologies:** translation, personal assistants, MDSM etc.
- **Artificial Intelligence:** Important breakthroughs and massive investments in R&D and applications (mostly in the US and Asia) – huge opportunity for Europe!
- **The European Language Challenge** cannot be abandoned or outsourced!

➢ **Need for Language Technology made *in* Europe *for* Europe!**

SRIA Version 1.0 beta

META NET

- SRIA V1.0 beta – unveiled at META-FORUM 2017 (13/14 Nov.)
- Prepared and presented by Cracking the Language Barrier federation
- Extended editorial team
- Document available on

  http://www.cracker-project.eu

  http://www.cracking-the-language-barrier.eu

Cracking the Language Barrier

Strategic Research and Innovation Agenda

Language Technologies for Multilingual Europe

Towards a Human Language Project

SRIA Editorial Team

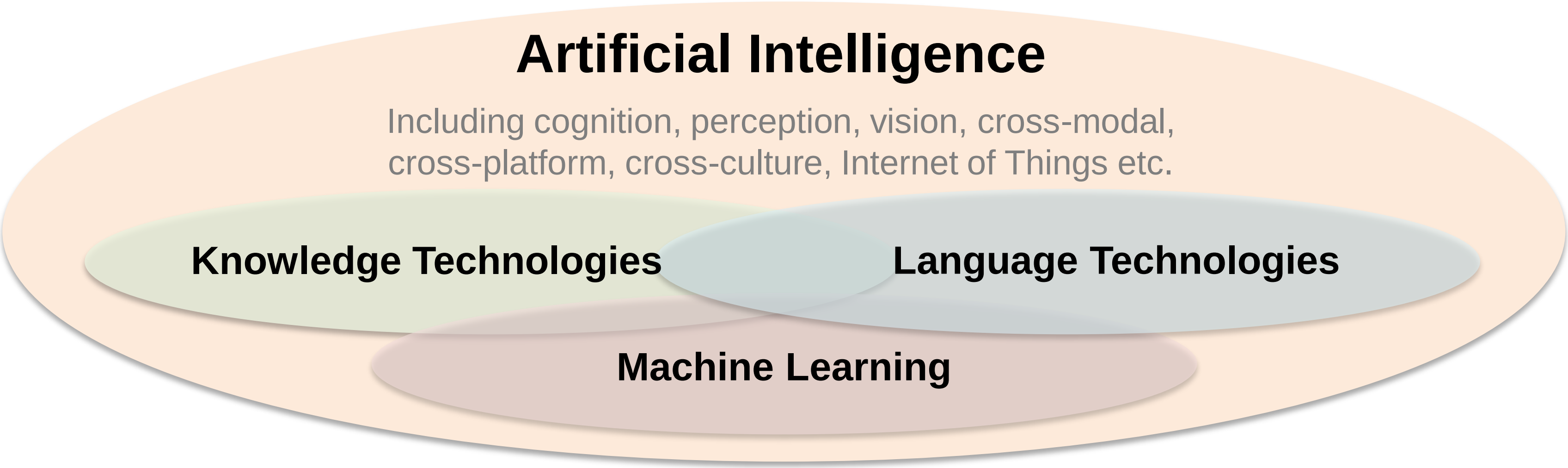
Version 1.0 beta – November 2017

Cracking the Language Barrier

Human Language Project

Human Language Project

- Large-scale EU funding programme – 10-15 years
- Goal: Deep Natural Language Understanding by 2030
- Artificial Intelligence for Next Generation Language Technology!
- New breakthroughs and groundbreaking results for industry, society, innovation, economy (multilingual digital single market).



Artificial Intelligence

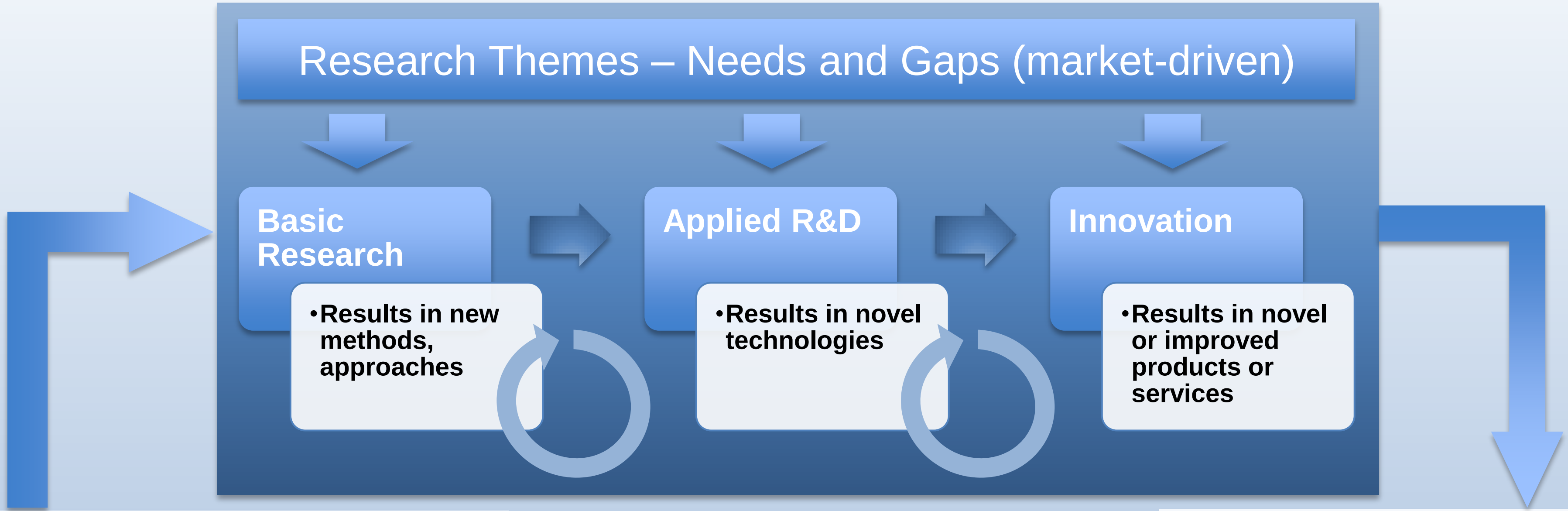
Including cognition, perception, vision, cross-modal, cross-platform, cross-culture, Internet of Things etc.

Knowledge Technologies

Language Technologies

Machine Learning

Human Language Project – Interdisciplinary R&D&I Programme



Research Themes – Needs and Gaps (market-driven)

Basic Research

- **Results in new methods, approaches**

Applied R&D

- **Results in novel technologies**

Innovation

- **Results in novel or improved products or services**

- Computational Linguistics
- Artificial Intelligence
- Language Technology
- Linguistics
- Computer Science
- Cognitive Science
- *other related fields*

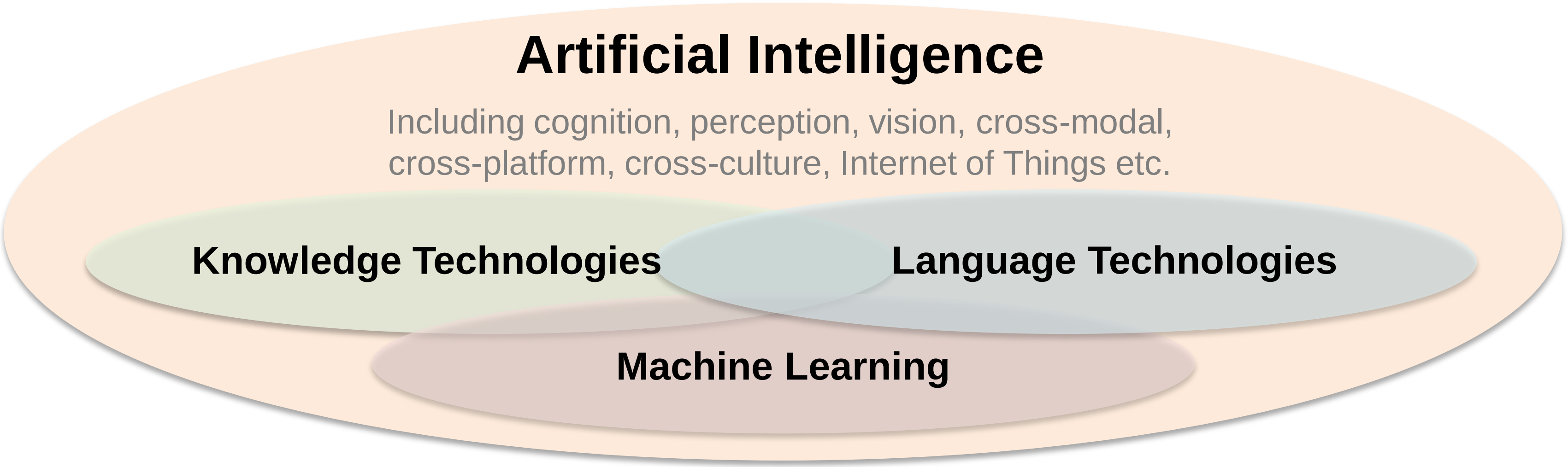
HLP: Umbrella programme to turbo-charge and to coordinate all European R&D&I activities in a systematic way including EP, EC, Member States.

- New, groundbreaking methods, paradigms, approaches
- Foster technologies, products, innovation, economy
- Foster education

Human Language Project

- **Goal: Deep Natural Language Understanding**
- **Breakthroughs in Artificial Intelligence plus a fresh look at Linguistics for the Next Generation of LT!**
- **All official European and many additional languages**
- **Broad coverage, high quality, high precision**
- **Create approaches, algorithms, data sets, resources**
- **Across modalities:** text, text types, speech, image, video etc.
- **Across platforms:** messaging, telephony, social, mobile, IoT etc.
- **Across cultures:** knowledge, customs, formalities, humour, emotion, subjectivity, biases, opinions, filter bubble etc.

Key Ingredients



Artificial Intelligence

Including cognition, perception, vision, cross-modal, cross-platform, cross-culture, Internet of Things etc.

Knowledge Technologies

Language Technologies

Machine Learning

- Extend knowledge bases
- Semantic Web, ontologies, linked data, interoperability
- More complex models
- Multilingual resources that are grounded, extensible
- Subjectivity, objectivity, further novel dimensions
- Web-scale reasoning

- Combine DNNs and symbolic processing
- ML for knowledge acquisition and extension
- DNNs embedded into modular systems including symbolic knowledge bases
- Make it possible to inspect and also to optimise DNNs (beyond end-to-end)

- (Computational) Linguistics research towards deep language understanding
- From corpora to DNNs to annotated data to highly improved symbolic methods
- Language portability
- *Full and Deep Language Understanding by 2030 –* ***Human Language Project***

HLP: Selected Topics

META NET

- **High-Quality MT** – overcome quality (and language) barriers, written and spoken, collaborate closely with human translators
- **Content Curation** and Smart Online Content
  - Increasing commercial and social relevance of content (“fake news”)
  - Include: domain, text type, style, register, discourse, social etc.
  - Type-specific, genre-specific analysis, assessment, generation
- **Multilingual European Knowledge Graph** that consolidates existing and emerging data (for crosslingual search, BI etc.)
- **Conversational interfaces**, especially for IoT, WoT, Industrie 4.0
- **Multilingual Europe**: LRs and LTs for *all* European languages
  - Include Member States – make it *coordinated*, *shared*, *focused*
  - Set of basic tools as open source and SaaS (free of charge)
  - Goal: boost the LT ecosystem and MDSM

Thank you. Questions?

office@meta-net.eu

http://www.meta-net.eu
http://www.facebook.com/META.Alliance