Engaging Content
Engaging People

Understanding the World:
Bridging Multilingual Knowledge and Translation

Dimitar Shterionov, *DCU / ADAPT*
Maite Melero, *UPF*
Marco Turchi, *FBK*
Dave Lewis, *TCD / ADAPT*

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
sfi Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Agenda

- Participants
- Data driven Machine Translation
- Multilingualism in Public Sector
- Multimodality, knowledge bases/knowledge graphs, ontologies
- Applications and development
- User-centric multilingual services

Participants

Dr. Maite Melero is a Computational Linguist. She participated in early rule-based approaches to Machine Translation, such as the Eurotra project, in the 1990's and the first Microsoft machine translation system, in the early 2000s.

Later she contributed to the definition of the Parole-Eagles standards for lexical encoding. She is now an adviser to the Spanish Administration for the impulse to Language technologies within the Administration, including compilation of linguistic resources and data and adoption of assisted translation tools.

Participants

Marco Turchi is the head of the Human Language Technology Machine Translation group at Fondazione Bruno Kessler (FBK).

He holds a Phd from the University of Siena, Italy, on “Computer Engineering, adaptive information processing”. Marco joined FBK in 2002 after working in world-level research labs (e.g. Yahoo, XRCE and University of Bristol).

His main research area is machine learning applied to machine translation, with particular interests in automatic post-editing and neural models for MT. He has co-authored about 120 scientific papers, contributed in several international and national projects such as Matecat, MMT and QT21. He organises the Conference on Machine Translation and shared task on automatic post-editing and is program chair of the International Workshop on Spoken Language Translation. He serves as reviewer of the most important conferences and journals in natural language processing.

Participants

Dr. Dave Lewis is an associate professor at Trinity College Dublin and the Director of the Knowledge and Data Engineering Group (KDEG). He is also a PI in CNGL Centre for Global Intelligent and a PI in the ADAPT Centre – the successor to CNGL.

His research interest is in Management Knowledge, and how it can be captured, modelled, analysed and exchanged in decentralised decision-making settings. He investigates the use of declarative knowledge techniques, including open data formats, ontological knowledge formats, service specifications and workflow models.

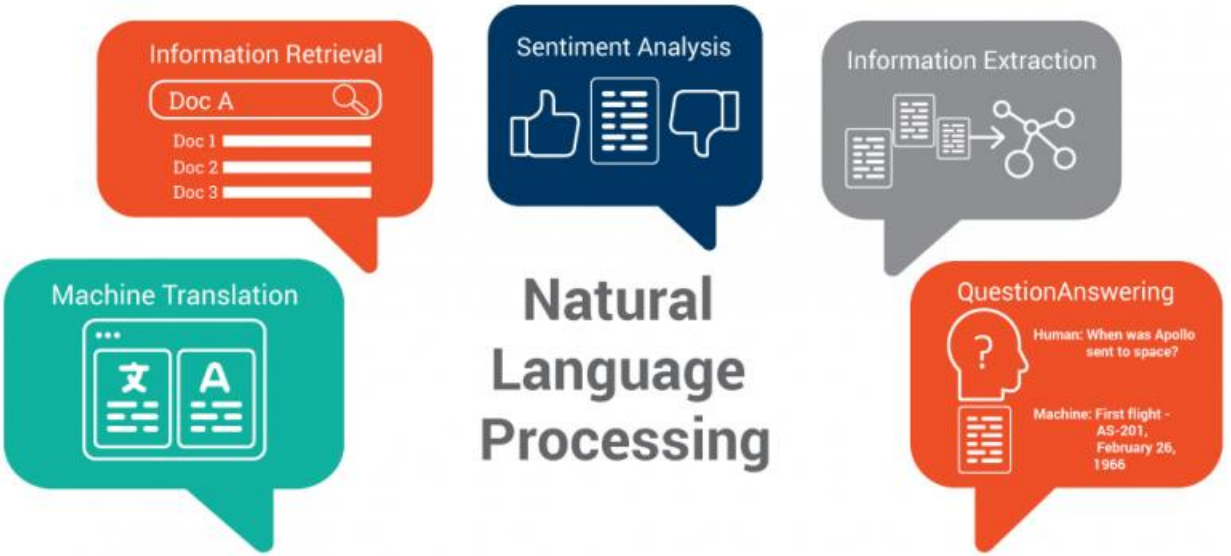
Application of these techniques and their evaluation has had impact in application areas such as Telecommunications Management, Smart Building Management, Online Communities, Intelligent Content Processing. He has pioneered the use of open semantic models to monitor the interplay between human language workers (e.g. translators & terminologists) and NLP components, and curating the result to improve those components.

Participants

Dimitar Shterionov is a post-doctoral researcher in ADAPT. He holds a PhD in computer science from KU Leuven Belgium. During his PhD, Dimitar has worked on the design and development of artificial intelligence software for learning and reasoning with uncertain data.

Previously he has led KantanLabs – a research and development group committed to advancing language technology within which he worked on introducing innovative technology such as efficient word reordering, improved alignment, neural MT, multilingual translation, automatic post-editing, etc. He is a member of the ADAPT team since November 2017. He has worked on different topics and projects, some of purely research nature and others, more applied projects, in collaboration with industry partners.

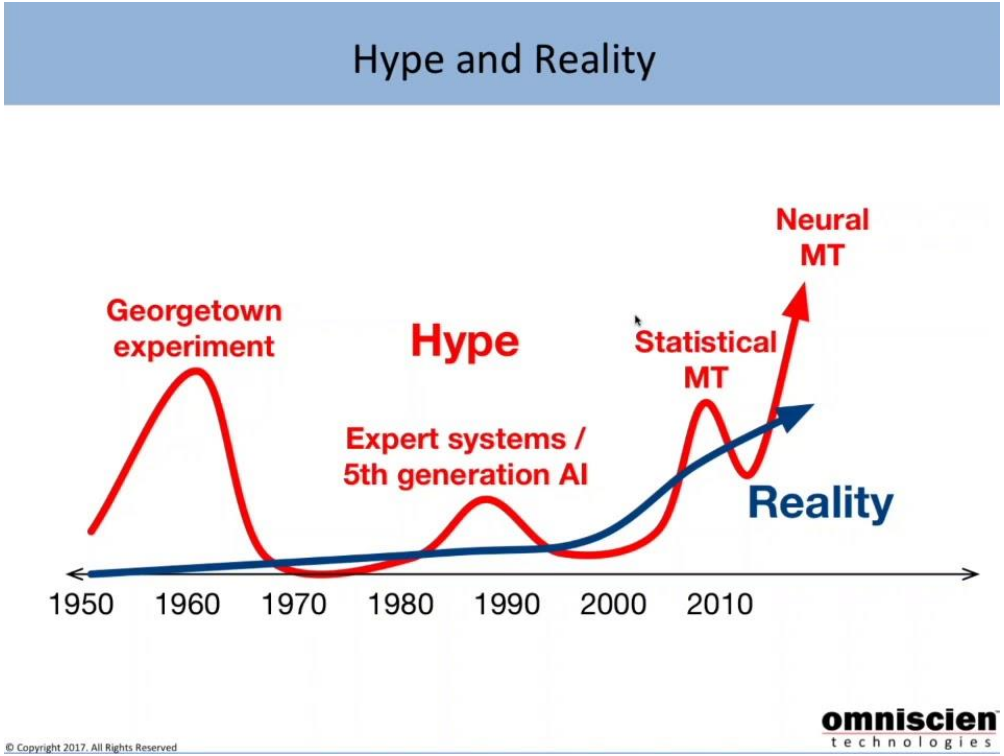
Problem of language



Problem of language

A mean to understand each other, a key to communication.

"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, 1947



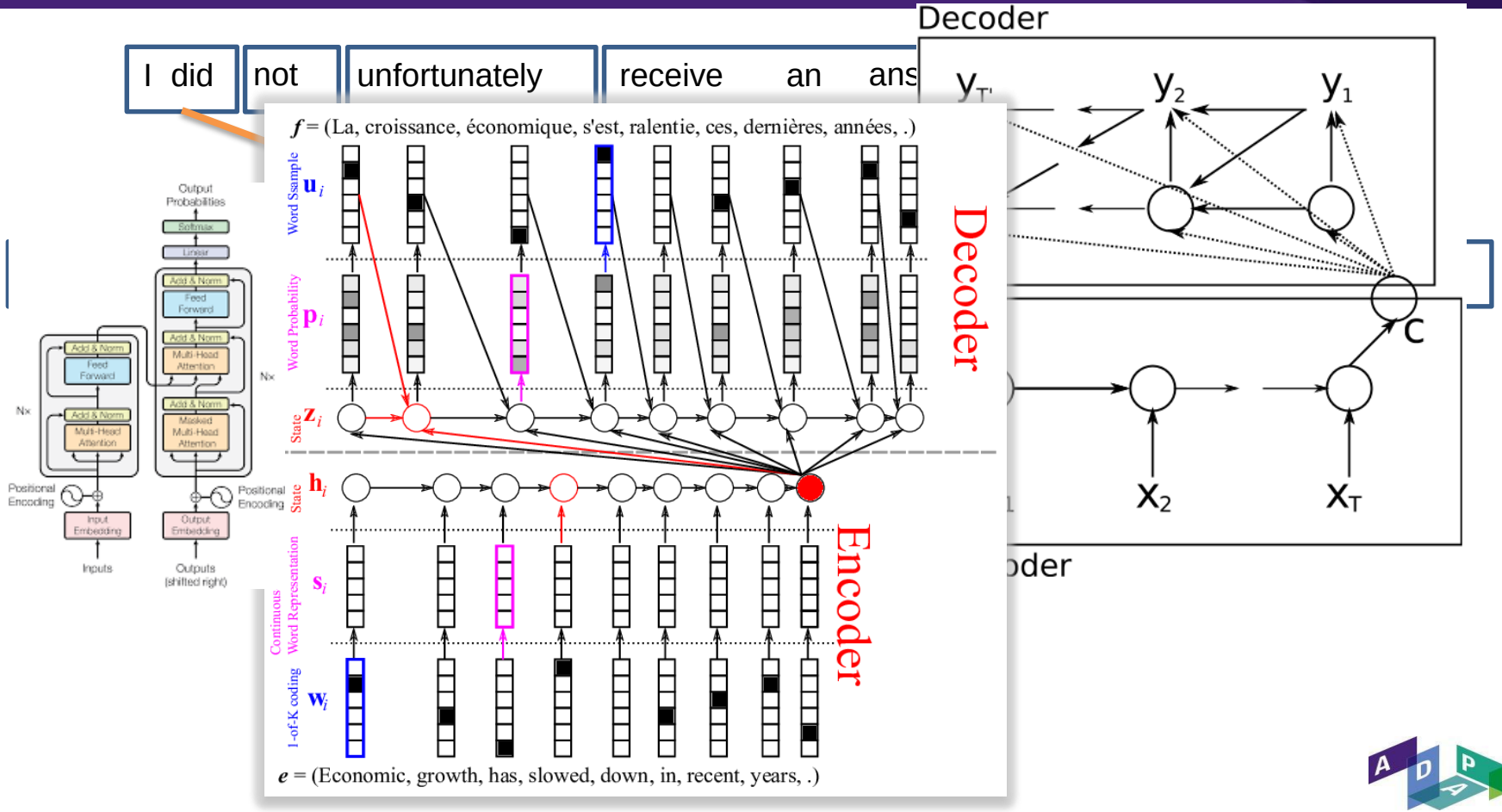
[Philipp Koehn, Webinar by Omniscien 2017.]

Problem of language

1. Translating out-of-domain data
2. The need for a lot of training data
3. Translating rare / unknown words
4. Handling of long sentences
5. No word alignments
6. Inconsistency at decoding time
7. Results are not very interpretable

[http://iconictranslation.com/2018/08/issue-4-six-challenges-in-neural-mt/]

A mean to understand each other, a key to communication.



Training data = parallel sentences

- Syntax and semantics derived from data
- Efficient training, retraining and translation
- Quality ~ quantity
- Good generalization ~ bad specialization

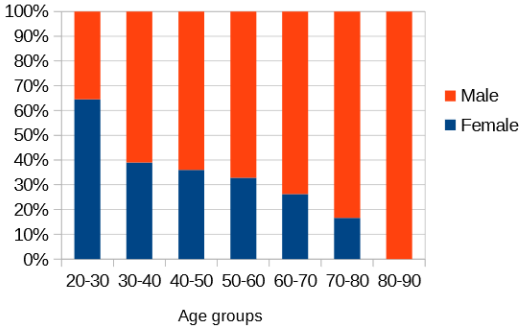
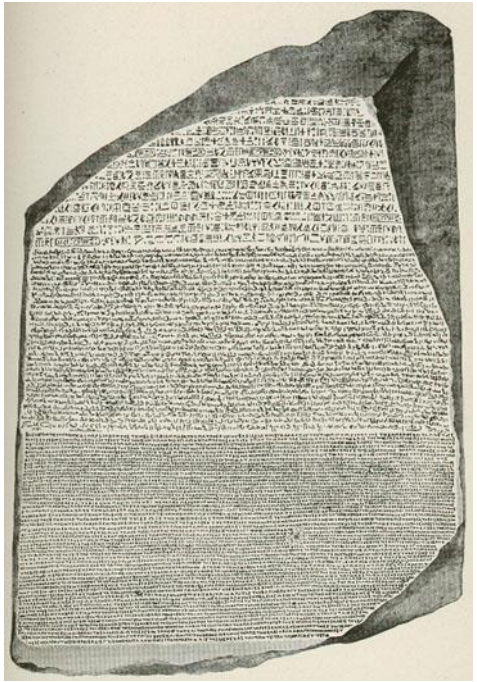


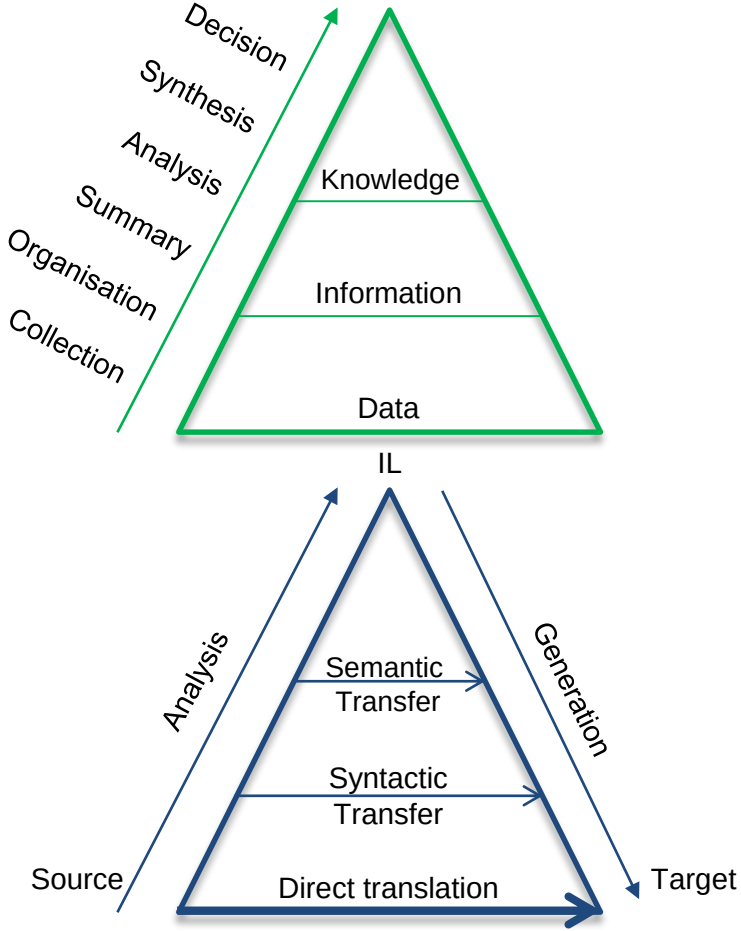
Figure 1: Percentage of female and male speakers per age group

(Ref)	En tant que *vice-président*...
(BASE)	En tant que *vice-présidente*...
(TAG)	En tant que *vice-président*...
(Ref)	... je suis *heureuse* que...
(BASE)	... je suis *heureux* que...
(TAG)	... je suis *heureuse* que...

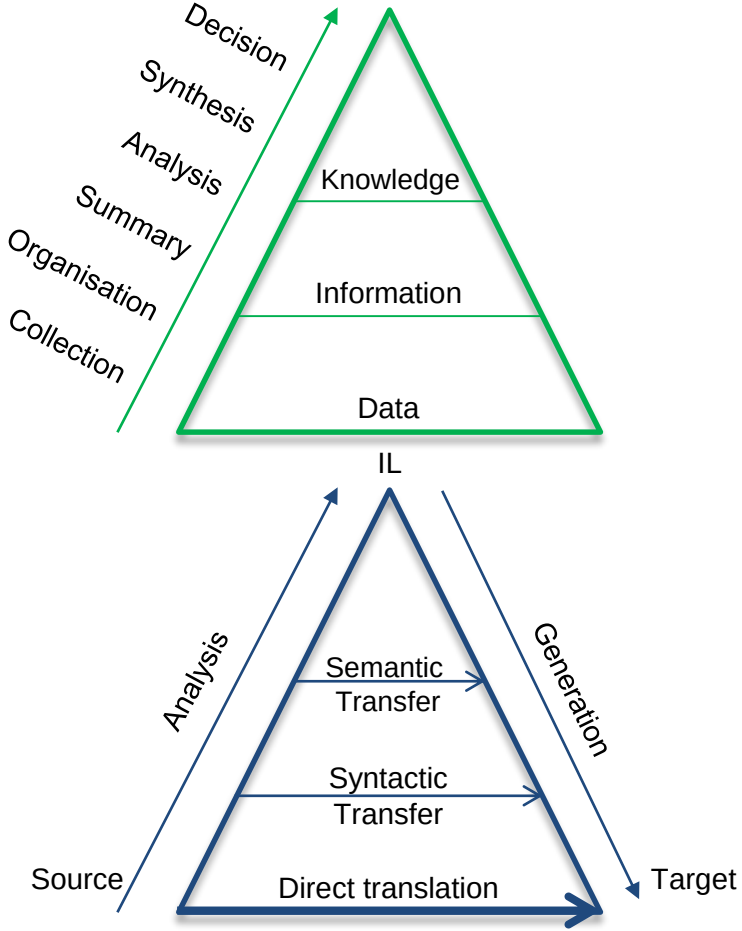


[Vanmassenhove et al, 2018, Getting Gender Right in Neural Machine Translation]

Data, knowledge, understanding



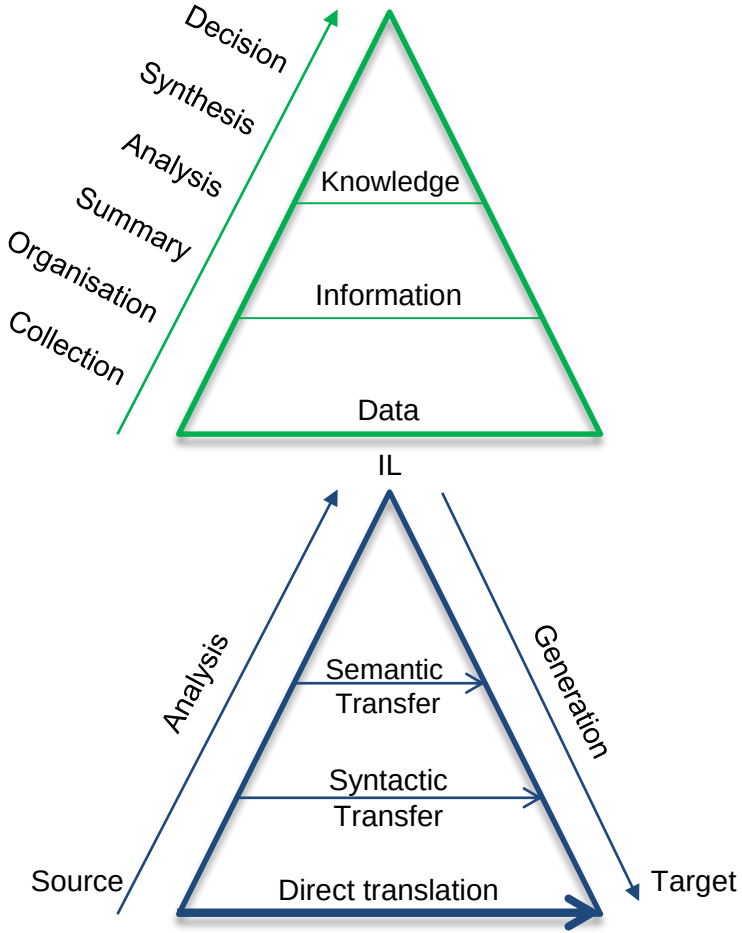
Data, knowledge, understanding



- Domain adaptation
- Discourse MT (context-aware MT)
- Interactive MT
- Constrained decoding

Textual context

Data, knowledge, understanding



Textual context

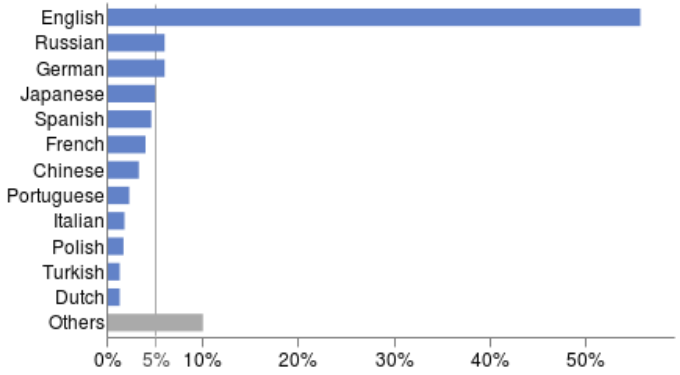
- Domain adaptation
- Discourse MT (context-aware MT)
- Interactive MT
- Constrained decoding

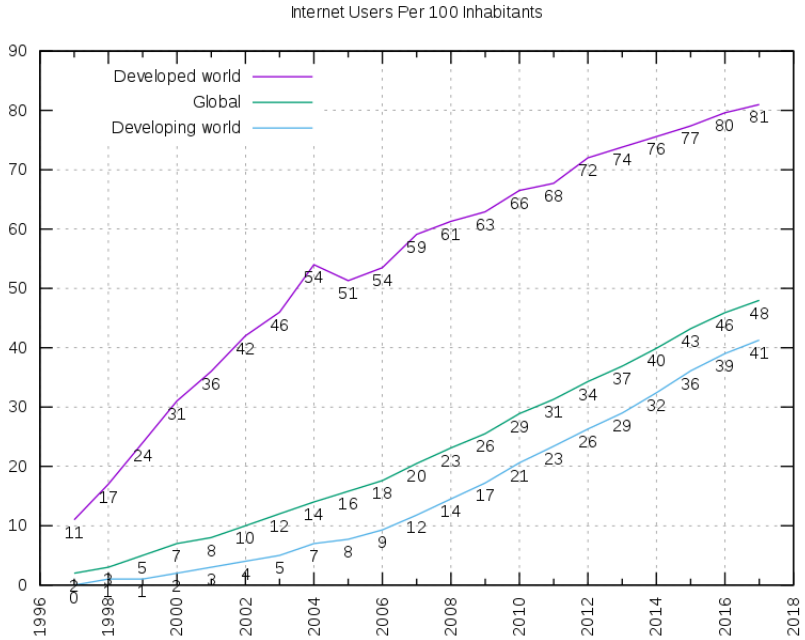
Multimodality

- Linguistically augmented MT
- Image + text MT
- Speech-to-text, speech-to-speech MT
- Automatic caption generation

Languages and users on the web

Country or area	Internet users	Rank	Percentage	Rank
China	746,662,194	1	53.20%	109
India	391,292,635	2	29.55%	143
United States	245,436,423	3	76.18%	54
Brazil	123,927,230	4	59.68%	90
Japan	117,528,631	5	92.00%	15
Russia	110,003,284	6	76.41%	53
Mexico	75,937,568	7	59.54%	92
Germany	73,436,503	8	89.65%	20
Indonesia	66,244,991	9	25.37%	157
United Kingdom	62,354,410	10	94.78%	11
Pakistan	62,000,000[8]	11	29.92%	143
Philippines	57,342,723	11	55.50%	101
France	55,413,854	12	85.62%	29
Nigeria	47,743,541	13	25.67%	153
South Korea	47,094,267	14	92.72%	14
Turkey	46,395,500	15	58.35%	95
Vietnam	43,974,618	16	46.50%	120





Multilingualism in EU

- EU delegates, international journalists, foreign visitors at the events organized by each hosting country of the EU Council Presidency represent speakers of (at least the) 24 official languages of the EU.
  [Pinnis and Kalnins, 2018, Developing a Neural Machine Translation Service for the 2017-2018 European Union Presidency]
- Government policy discussions often involve extensive research, and compilation of comparative studies or trend data collected from several countries across the EU or even the world. There has been an increasing need for central and local government to translate and incorporate multilingual content into documentation and this is also important when it comes to face-to-face meetings or conferences, in which an interpreter may be required.
  [https://www.tjc-global.com/government_translation_translators_worldwide/]

- *English, please?*

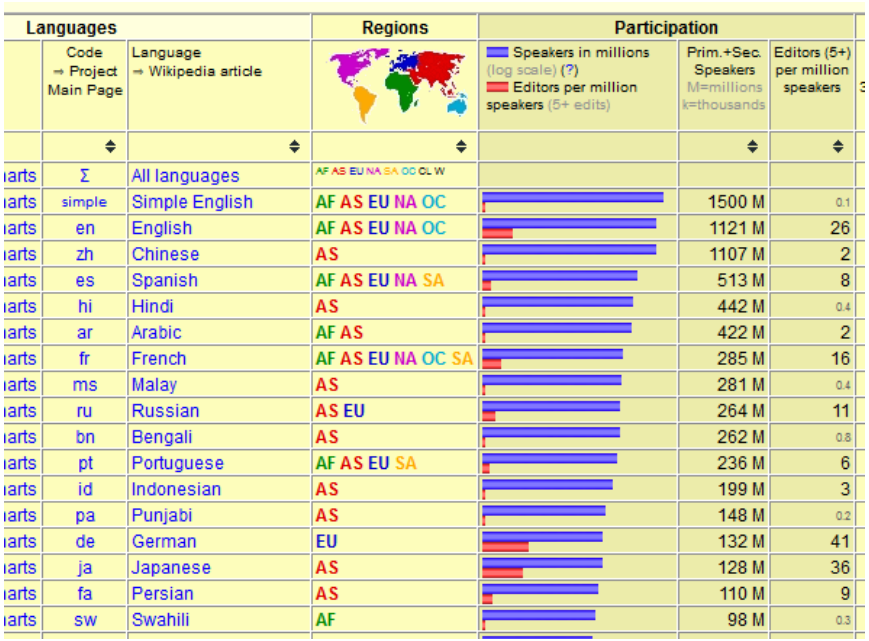
A multilingual expanse

With the constant flow of people across the borders of EU member states it is crucial that the public services provide multilingual support.

- How much multilingual support is required in the Public Administration? Isn't google translate enough? What do you imagine needs to be done in order to provide the necessary language technology for smooth communication that leads to unproblematic integration?
- What is your opinion on English as a lingua franca vs services in multiple languages?
- How about multilingual MT, e.g., transfer learning? zero-shot translation? pivot-based MT?

Multimodal content

“Over 400 hours of content are uploaded to YouTube every minute, which equates to a staggering 576,000 hours of content per day, likely making the task of content moderation arduous.”
[https://www.businessinsider.de]



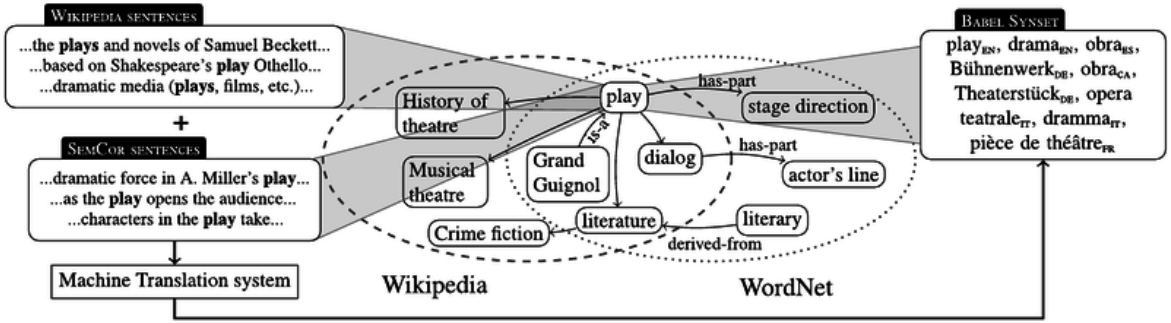
Languages			Regions	Participation		
	Code ⇒ Project Main Page	Language ⇒ Wikipedia article		Speakers in millions (log scale) (?) Editors per million speakers (5+ edits)	Prim.+Sec. Speakers M=millions k=thousands	Editors (5+) per million speakers
arts	Σ	All languages	AF AS EU NA SA OC CL W			
arts	simple	Simple English	AF AS EU NA OC		1500 M	0.1
arts	en	English	AF AS EU NA OC		1121 M	26
arts	zh	Chinese	AS		1107 M	2
arts	es	Spanish	AF AS EU NA SA		513 M	8
arts	hi	Hindi	AS		442 M	0.4
arts	ar	Arabic	AF AS		422 M	2
arts	fr	French	AF AS EU NA OC SA		285 M	16
arts	ms	Malay	AS		281 M	0.4
arts	ru	Russian	AS EU		264 M	11
arts	bn	Bengali	AS		262 M	0.8
arts	pt	Portuguese	AF AS EU SA		236 M	6
arts	id	Indonesian	AS		199 M	3
arts	pa	Punjabi	AS		148 M	0.2
arts	de	German	EU		132 M	41
arts	ja	Japanese	AS		128 M	36
arts	fa	Persian	AS		110 M	9
arts	sw	Swahili	AF		98 M	0.3

“Twitter Demographics
24% of All Internet male users use Twitter, whereas 21% of All Internet Female users use Twitter.
There are 261 million International Twitter users which account for 79% of Twitter accounts are based outside the United States.
There are over 69 million Twitter users in the US.
Roughly 46% of Twitter users are on the platform daily.
The total number of Twitter users in the UK is 13 million.
37% of Twitter users are between the ages of 18 and 29, 25% users are 30-49 years old.
56% of Twitter users $50,000 and more in a year.
The top three countries by user count outside the U.S. are Brazil (27.7 million users), Japan (25.9 million), and Mexico (23.5 million).
36% of Americans aged 18 to 29 years old use Twitter.
80% of Twitter users accessing the platform on a mobile device, and 93% of video views are on mobile.”
[https://www.omnicoreagency.com/twitter-statistics/]

Multimodal content

DNNs are ~~becoming~~ the standard ML tool. In order to benefit from the advantages of DL over other ML techniques it is crucial to provide the right training. Combining different types of data - text, audio, image - can lead to a better DL model for a given task. These different types of data are typically very strongly related, e.g., an image represents context.

- What can be learned from and what are the advantages of structured data (knowledge graphs/bases, ontologies) in relation to MT systems?
- How important is the structure and the dependencies between the different types of data?
- Can MT be applied on ontologies, knowledge bases/graphs?



The engineering side

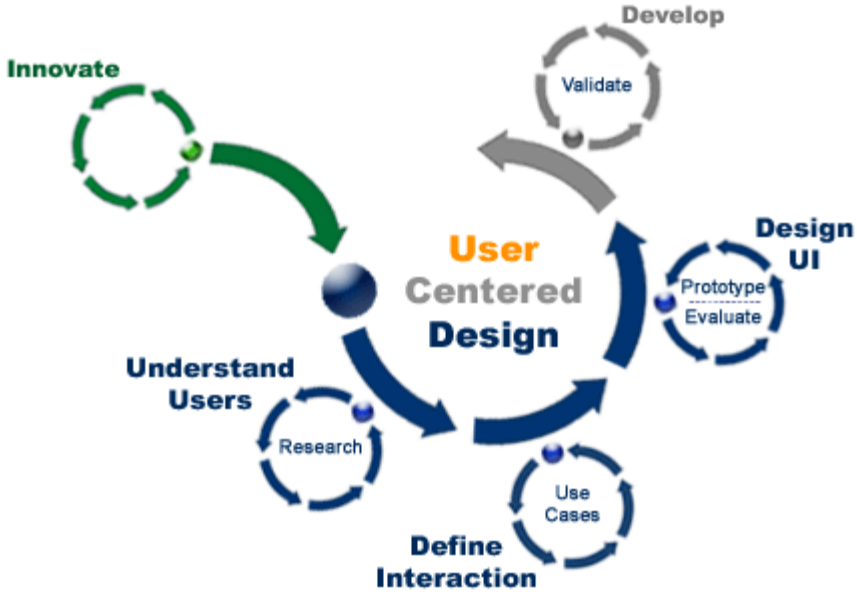
Multilingual services can be implemented as either a collection of multiple components (e.g, microservices) or as an end-to-end, unified system.

- How would you approach a multilingual problem such as, e.g., a multilingual chatbot?
- What symbiosis between application and language service providers do you see as the most viable? For example integrated MT system within the application itself vs. distributed/microservice architecture?

Who drives multilingual development onwards?

After all, what matters is the user's experience. While talking about understanding the world we may mostly think of (machine) translation -- converting data from one language to another -- we need to also bring the user as an architect of the app that perfectly suits their needs. That requires the active interaction and learning from user's experience.

- How do we employ the user's experience to improve the user's experience?



Make data, information, knowledge accessible