eTranslation as a EU Flagship Use Case of Natural Language Processing

SLOVKO 2023
Natural Language Processing and Corpus Linguistics

18-20 October 2023
Bratislava, SK

Directorate General for Translation
Daniel Kluvanec
Adviser

Directorate General for Translation (DGT) Artificial Intelligence + eTranslation projects

Jörgen Gren
Director resources

Dieter Rummel
Head of IT

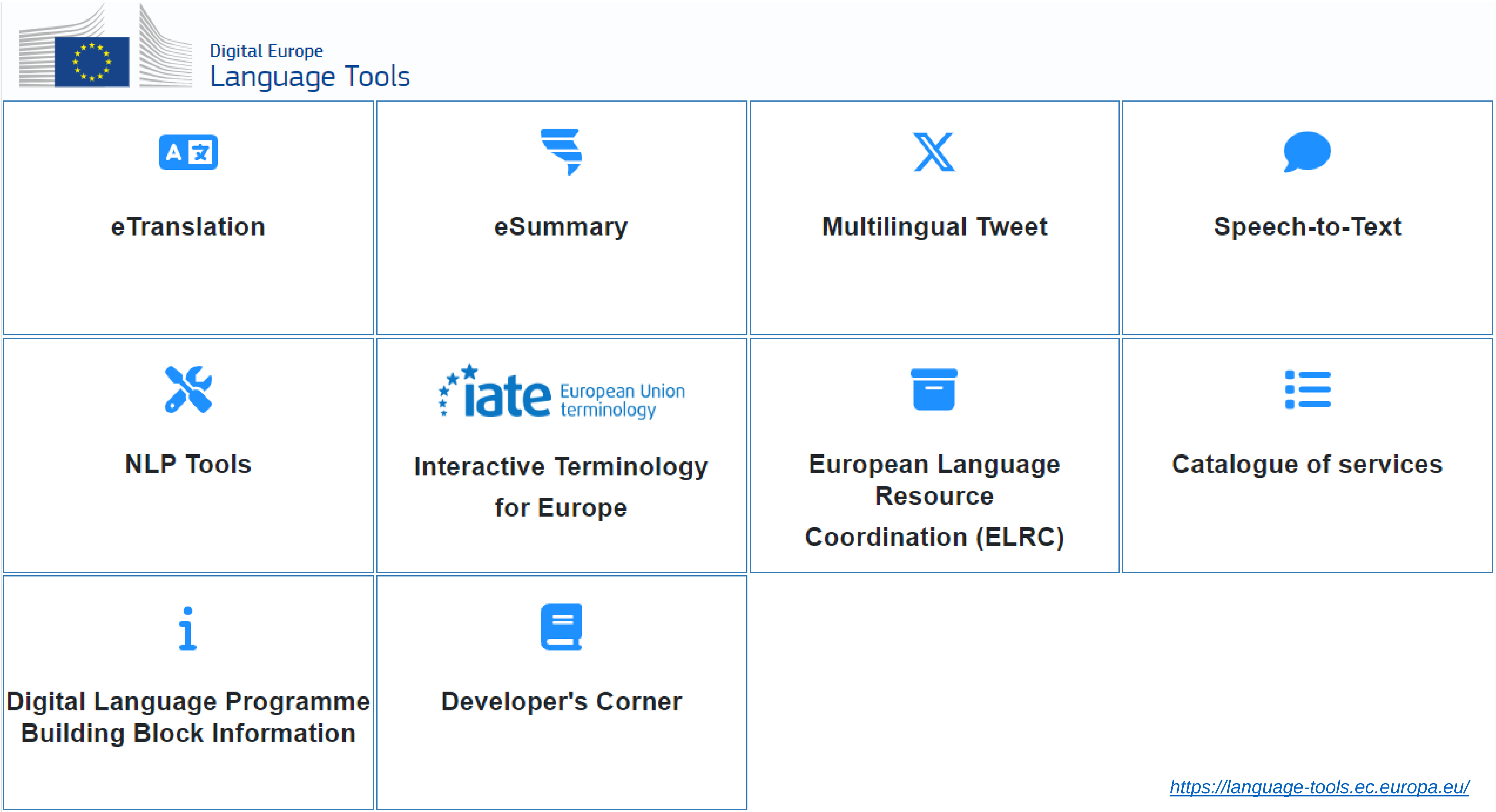
- eTranslation
- Language Tools

In support:

- Horizon Europe
- Digital Europe Programme

Markus Foti
eTranslation

Cristian
Brasoveanu
AI@EC



Digital Europe
Language Tools

eTranslation

eSummary

Multilingual Tweet

Speech-to-Text

NLP Tools

iate European Union terminology

Interactive Terminology for Europe

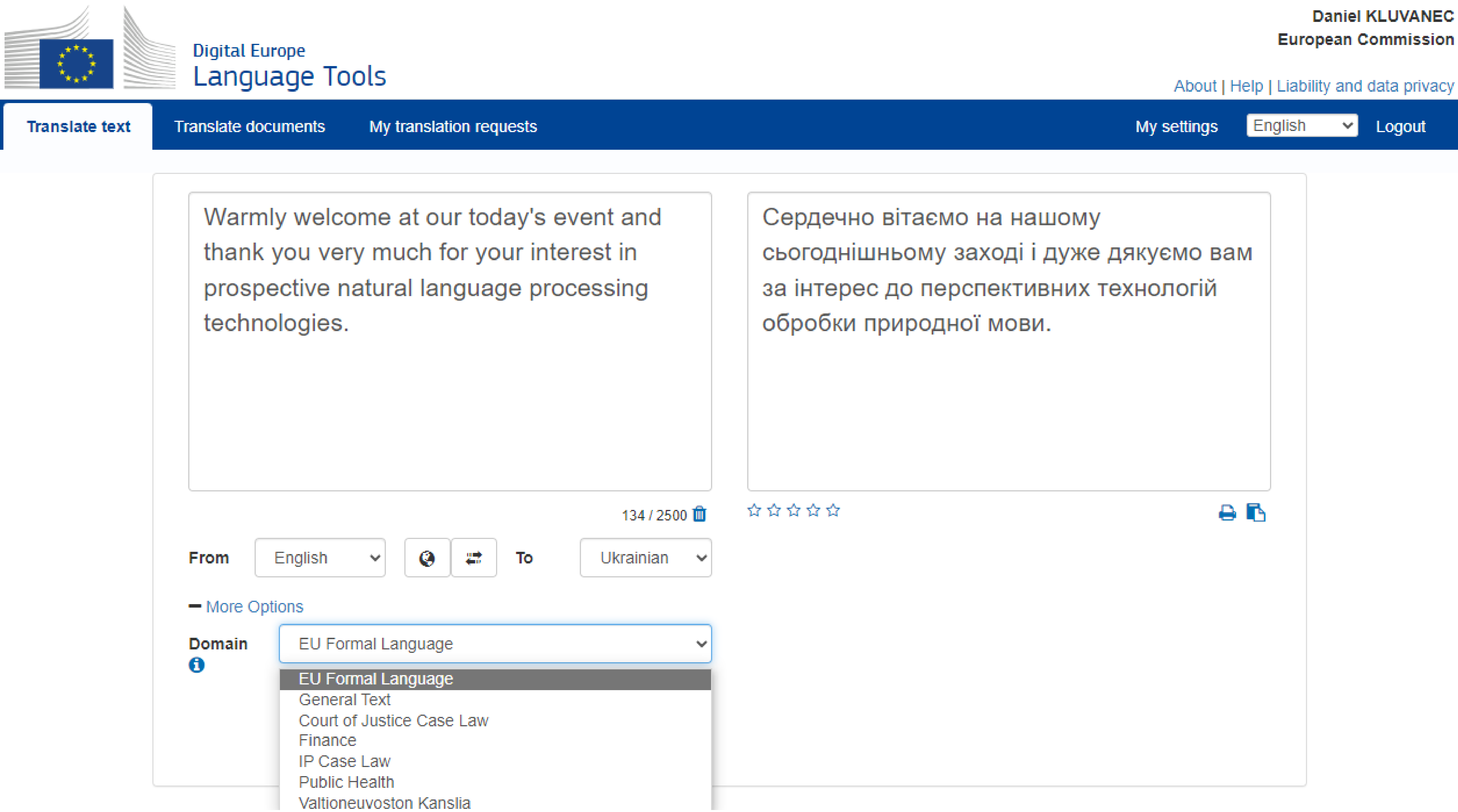
European Language Resource Coordination (ELRC)

Catalogue of services

Digital Language Programme Building Block Information

Developer's Corner

https://language-tools.ec.europa.eu/

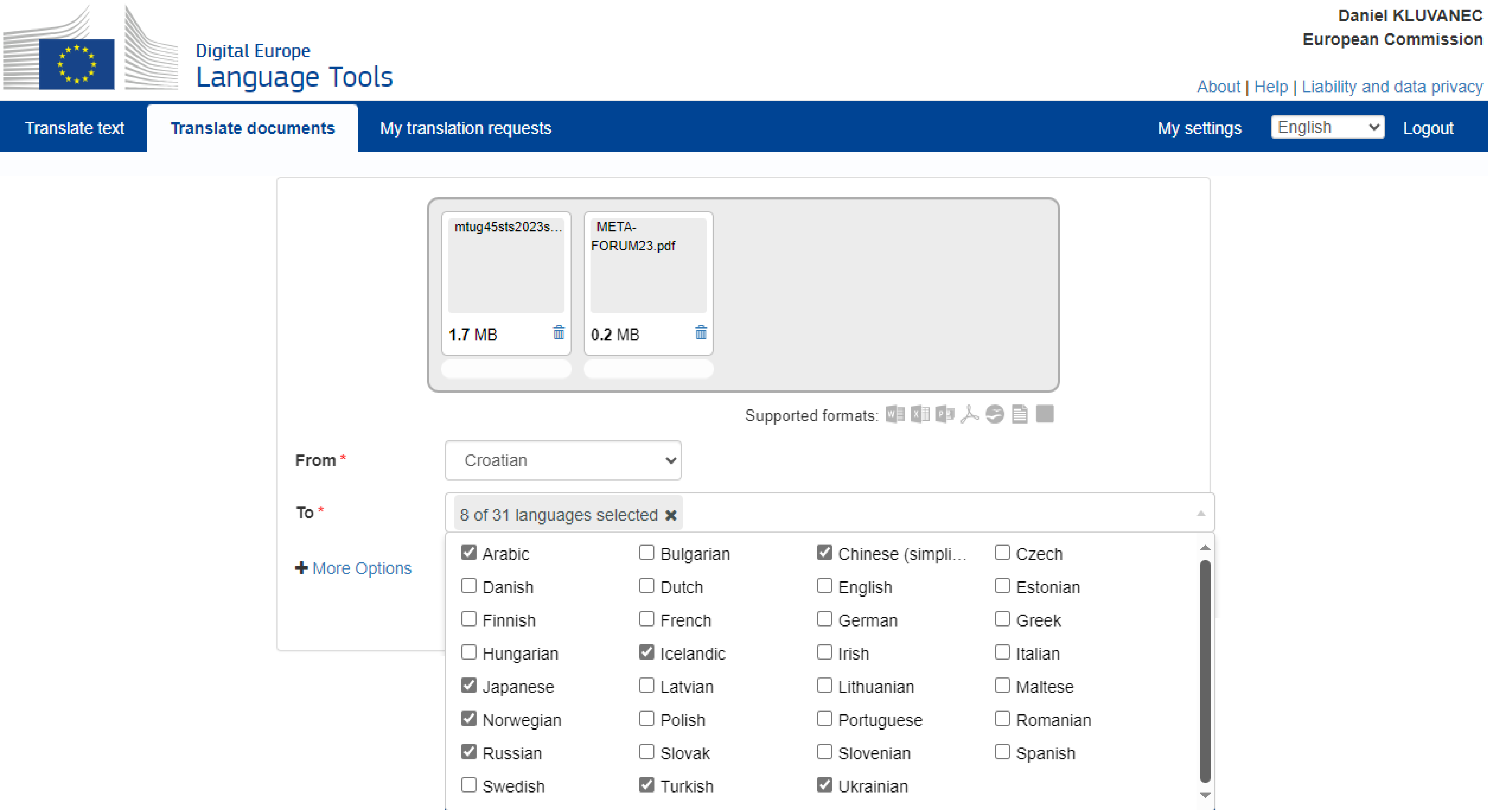


Warmly welcome at our today's event and thank you very much for your interest in prospective natural language processing technologies.

Сердечно вітаємо на нашому сьогоднішньому заході і дуже дякуємо вам за інтерес до перспективних технологій обробки природної мови.

From English **To** Ukrainian

Domain EU Formal Language



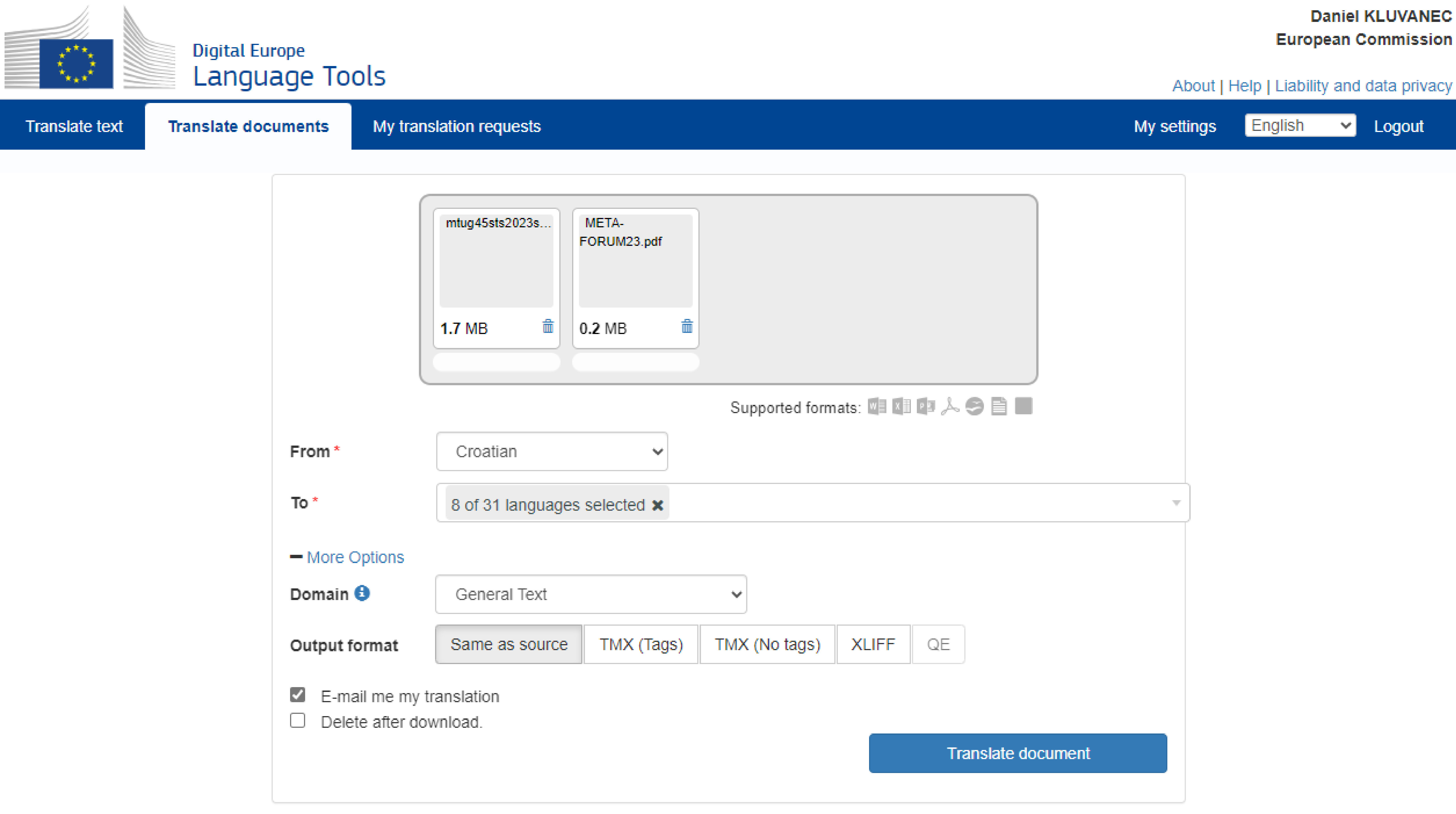
mtug45sts2023s... 1.7 MB

META-FORUM23.pdf 0.2 MB

From * Croatian

To * 8 of 31 languages selected ✖

☑ Arabic
☐ Bulgarian
☑ Chinese (simpli…
☐ Czech
☐ Danish
☐ Dutch
☐ English
☐ Estonian
☐ Finnish
☐ French
☐ German
☐ Greek
☐ Hungarian
☑ Icelandic
☐ Irish
☐ Italian
☑ Japanese
☐ Latvian
☐ Lithuanian
☐ Maltese
☑ Norwegian
☐ Polish
☐ Portuguese
☐ Romanian
☑ Russian
☐ Slovak
☐ Slovenian
☐ Spanish
☐ Swedish
☑ Turkish
☑ Ukrainian



From * Croatian

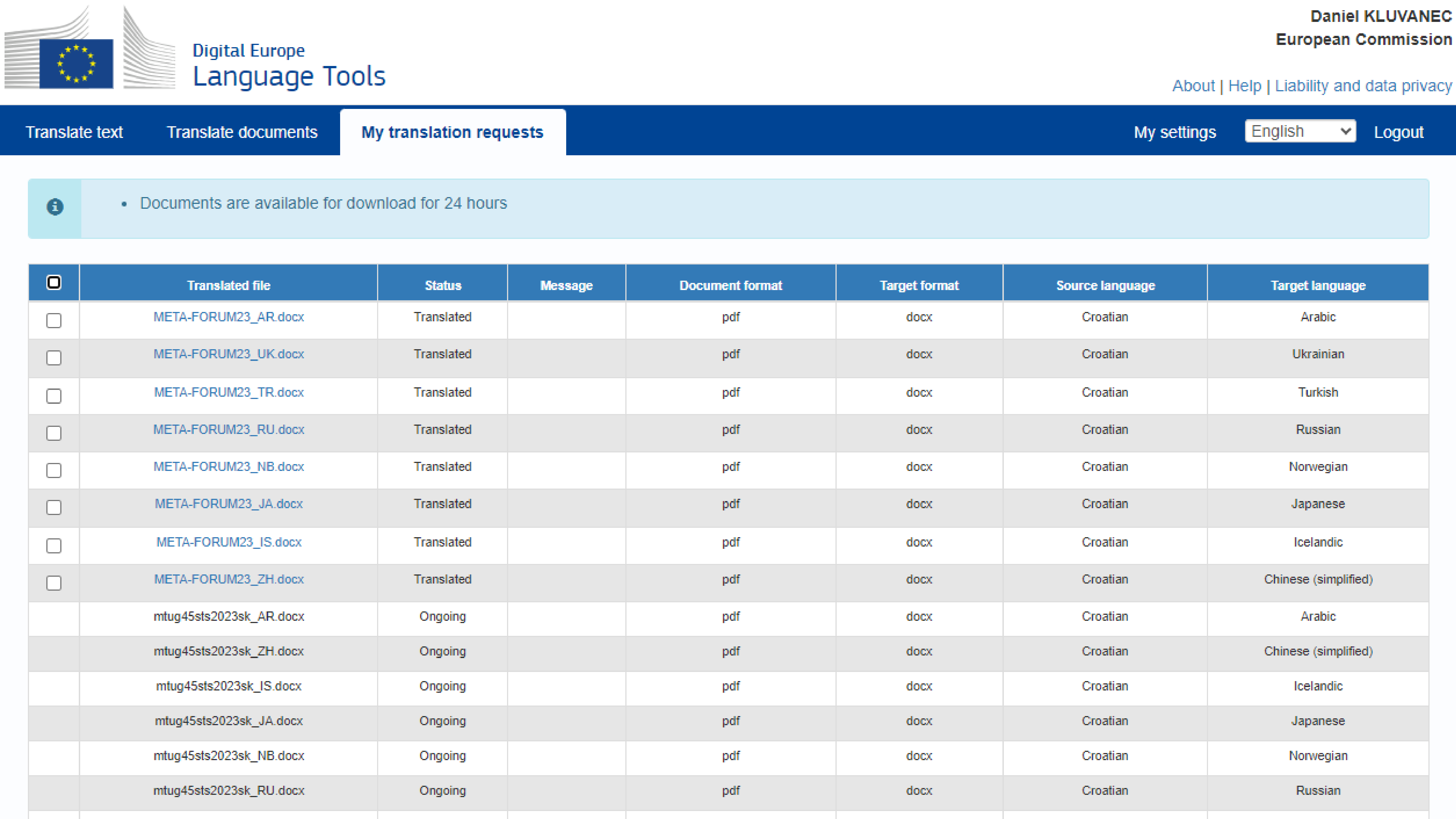
To * 8 of 31 languages selected ✖

Domain ℹ General Text

Output format Same as source

☑ E-mail me my translation

☐ Delete after download.



- Documents are available for download for 24 hours

☐	Translated file	Status	Message	Document format	Target format	Source language	Target language
☐	META-FORUM23_AR.docx	Translated		pdf	docx	Croatian	Arabic
☐	META-FORUM23_UK.docx	Translated		pdf	docx	Croatian	Ukrainian
☐	META-FORUM23_TR.docx	Translated		pdf	docx	Croatian	Turkish
☐	META-FORUM23_RU.docx	Translated		pdf	docx	Croatian	Russian
☐	META-FORUM23_NB.docx	Translated		pdf	docx	Croatian	Norwegian
☐	META-FORUM23_JA.docx	Translated		pdf	docx	Croatian	Japanese
☐	META-FORUM23_IS.docx	Translated		pdf	docx	Croatian	Icelandic
☐	META-FORUM23_ZH.docx	Translated		pdf	docx	Croatian	Chinese (simplified)
	mtug45sts2023sk_AR.docx	Ongoing		pdf	docx	Croatian	Arabic
	mtug45sts2023sk_ZH.docx	Ongoing		pdf	docx	Croatian	Chinese (simplified)
	mtug45sts2023sk_IS.docx	Ongoing		pdf	docx	Croatian	Icelandic
	mtug45sts2023sk_JA.docx	Ongoing		pdf	docx	Croatian	Japanese
	mtug45sts2023sk_NB.docx	Ongoing		pdf	docx	Croatian	Norwegian
	mtug45sts2023sk_RU.docx	Ongoing		pdf	docx	Croatian	Russian

Since 2012: plea for further research... vs. 2023

- Lexis?
- Syntax?
- Morphology?
- Semantics?
- Pragmatics?
- Reference check?
- Terminology consistency?
- Human skills transferable?
- New human interfaces, touch navigation?
- Seamless integration and CATT interoperability?
- Speedier / interactive machine learning cycles?
- Standard / cheaper hardware (GPGPU) reusability?
- Machine learning towards machine understanding?
- A merger with neurolinguistics and neuroinformatics?

Deep Learning (DL)
High Performance Computing (HPC)
Large Language Models (LLM)
Neural Machine Translation (NMT)

Guardians of Treaties: intentionality check – reserved for humans!

European Commission

Sources: Kluvanec et al. TEF / MT@Work 2012-2016, Tralogy 2013, STOA 2013, FIT 2014, JIAMCATT 2016, TIC 2018, Slovko 2023

The future of NMT & AI at DGT

MT@EC / eTranslation production for public administration use:

- fit for purpose in many administrative tasks
- in 2015 exceeded the DGT human-production (> 2 mio)
- in 2016 proven marginal / zero gap to LPE (Light Post-Editing) MT
- in 2018 more than 10-times of the DGT human production and full NMT coverage
- in 2022 almost 100-times of the DGT human production

Intentionality check = reserved for humans:

- today: 4 eyes principle / translator + reviser
- with narrow AI: 4 eyes principle / post-editor + reviser
- with general AI: ???
- (machine understanding + linguistic auditor...?)
- (automated support for more cognitive tasks and processes)

Source: Verleysen et al. CIUTI 2017

The Future of Employment 2013... vs. 2023

How susceptible are jobs to computerisation?

702 jobs ranked by probability of automation:

(excerpt)

- 0.6% Human Resources Managers
- 22% Statisticians
- 38% Interpreters and Translators
- 48% Computer Programmers
- 94% Accountants, Auditors, Budget Analysts
- 96% Secretaries and Administrative Assistants
- 98% Procurement Clerks

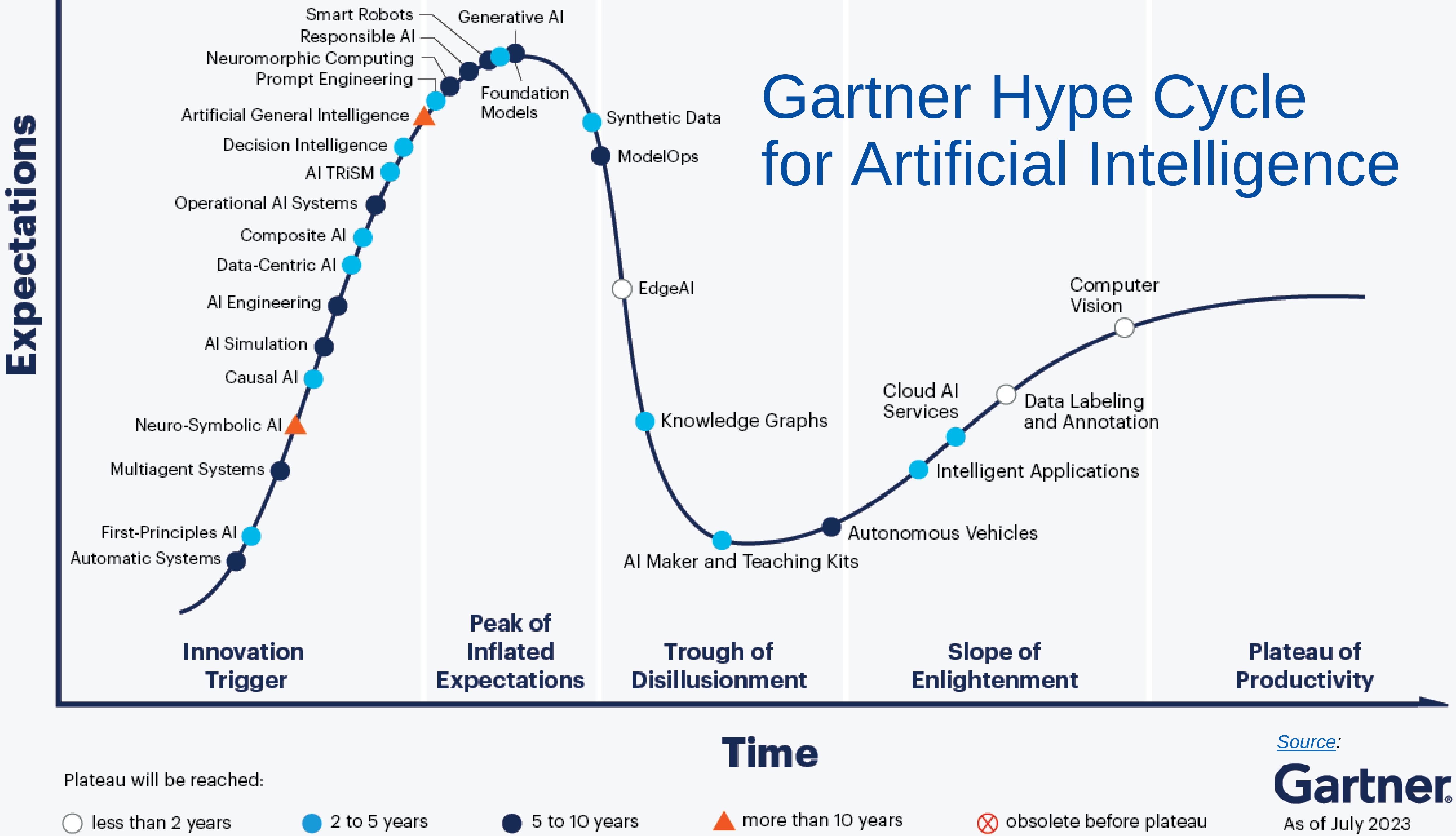
Source: Frey & Osborne, University of Oxford, 2013

Generative AI and the Future of Work: A Reappraisal (2023)

...the potential scope of automation is unlikely to substantially grow merely through scaling existing models. In conclusion, while we expect AI to continue to surprise us, and for many jobs to be automated away, in the absence of major breakthroughs, we also expect the bottlenecks we outlined in our 2013 paper to continue to constrain our automation possibilities for the foreseeable future.

Source: Frey & Osborne, University of Oxford, 2023

Gartner Hype Cycle for Artificial Intelligence



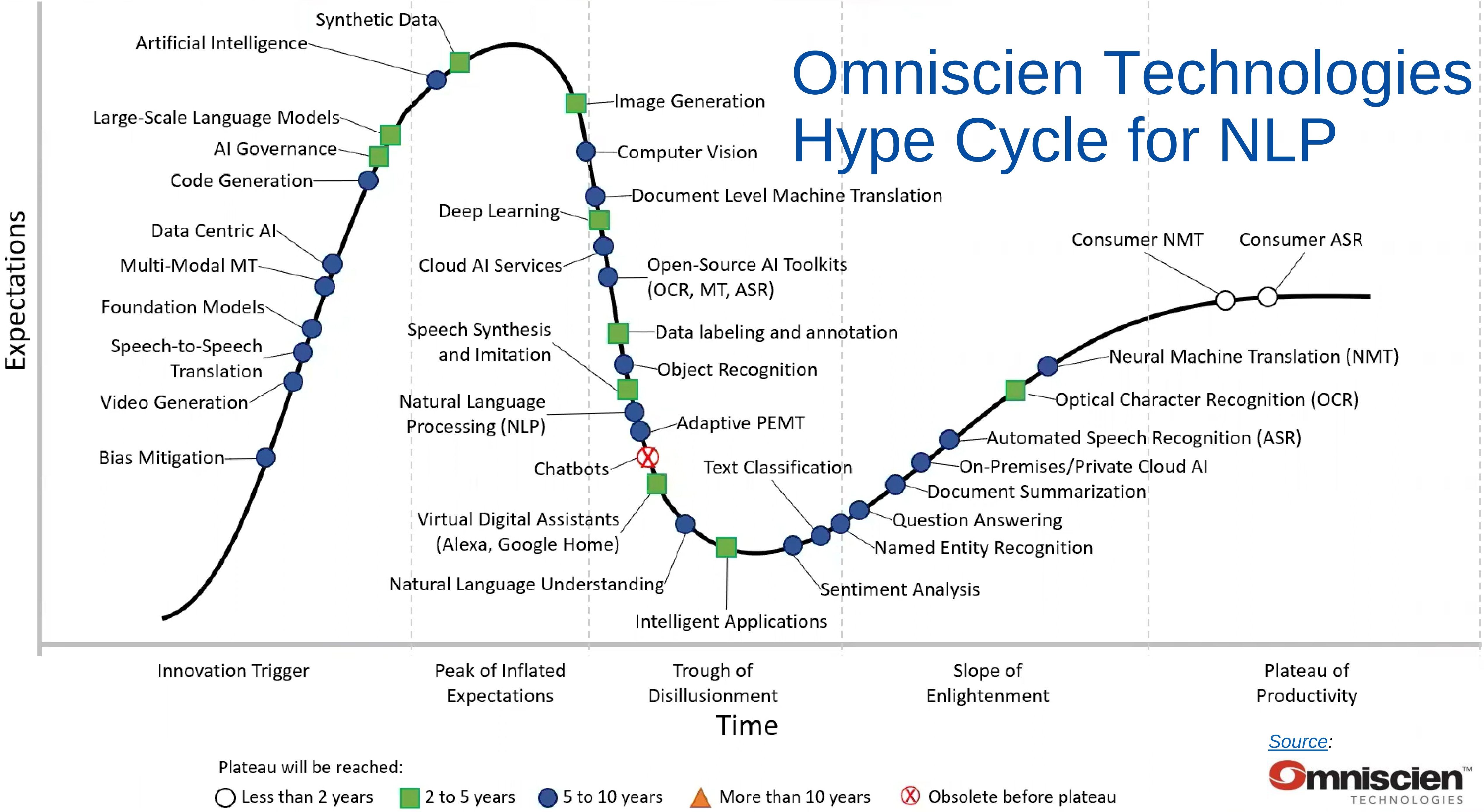
Source:

Gartner

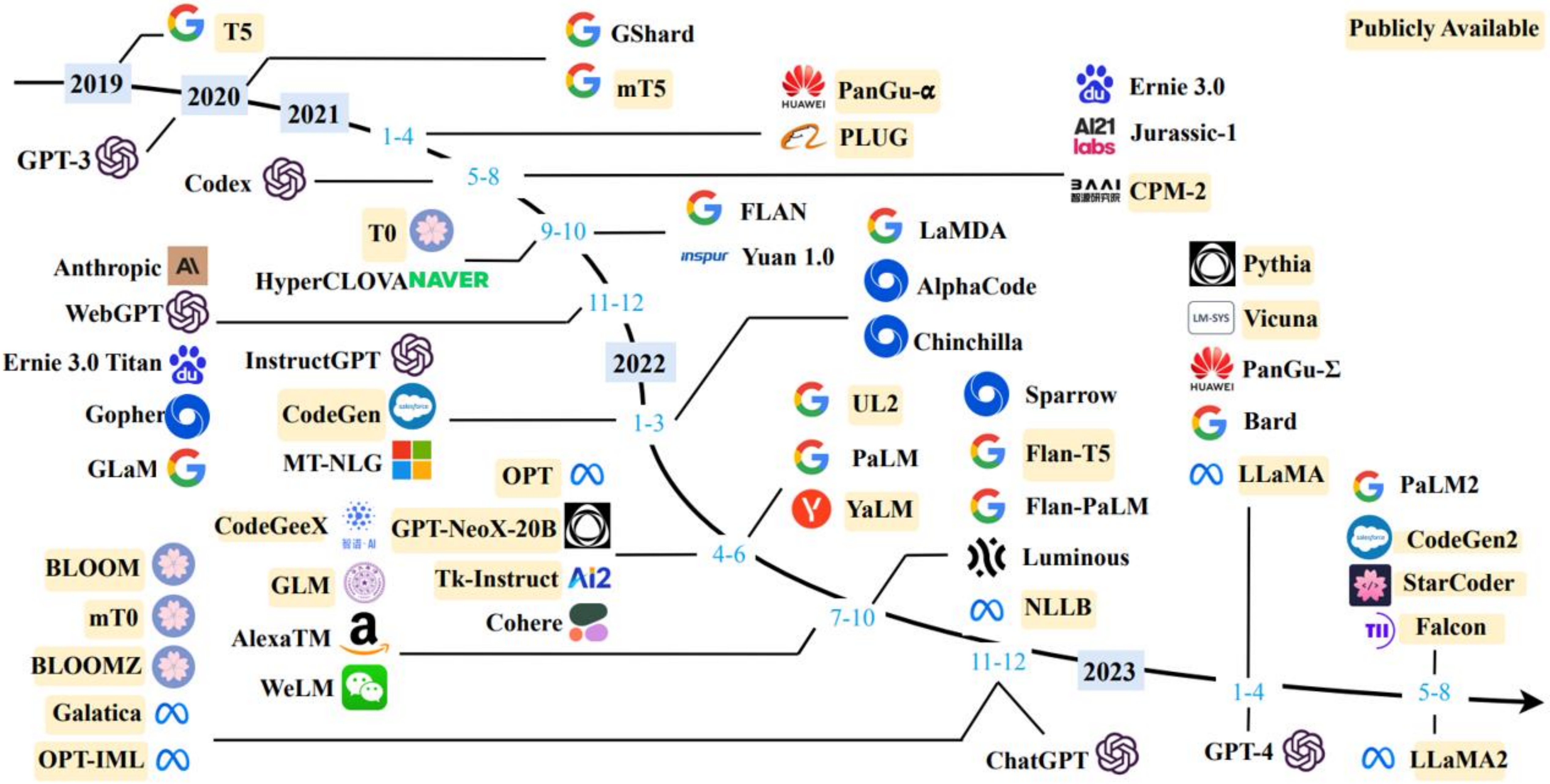
As of July 2023

January 2023

Omniscien Technologies Hype Cycle for NLP



Source:



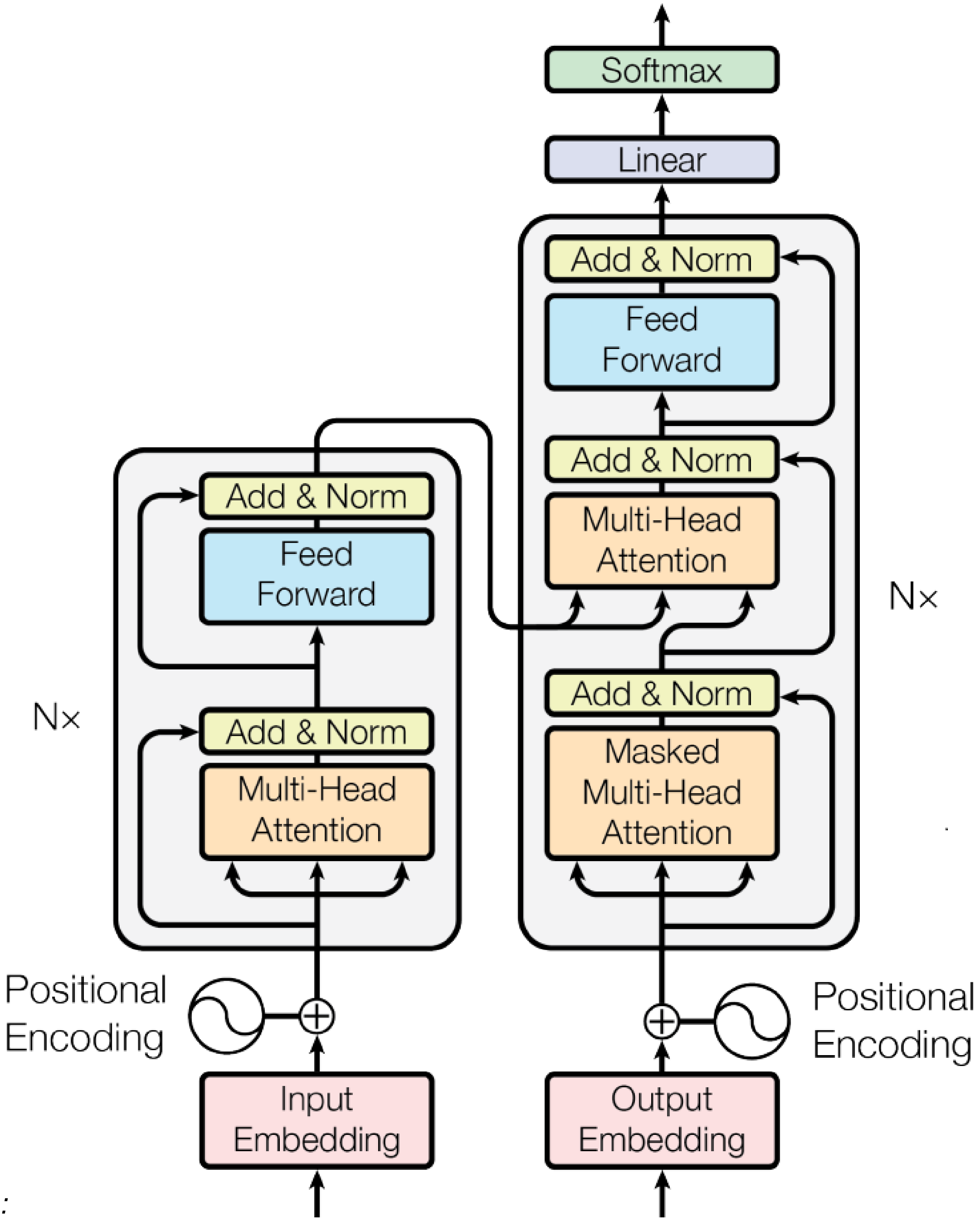
Source: Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie and Ji-Rong Wen, A Survey of Large Language Models, arXiv preprint, 2023

Transformer architecture

GPT = Generative Pre-trained Transformer
LLM = Large Language Model

- generates next word in a sequence
- averages text patterns from internet
- makes impression of a thinking machine
- can pass the Turing test
- can pass different encyclopaedic tests
- can reuse argumentation logic patterns

- cannot create a new inference
- hallucinates missing facts

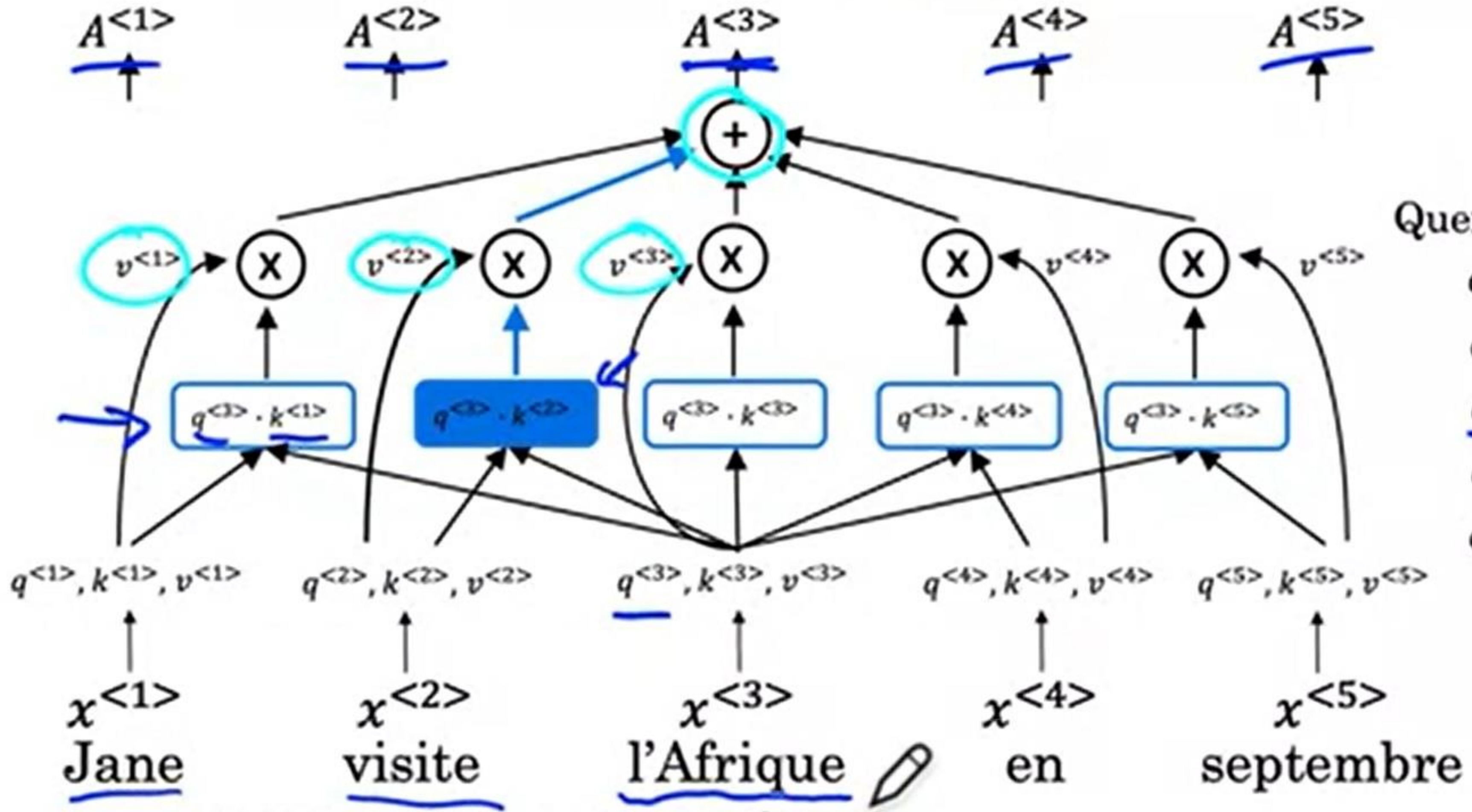


Source: Ashish Vaswani et aliiAshish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin: Attention Is All You Need, NIPS 2017

Self-Attention

$$A(q,K,V)=\sum_i \frac{\exp(q\cdot k^{<i>})}{\sum_j \exp(q\cdot k^{<j>})} v^{<i>}$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Query (Q)	Key (K)	Value (V)
$q^{<1>}$	$k^{<1>}$ person	$v^{<1>}$
$q^{<2>}$	$k^{<2>}$ action	$v^{<2>}$
$q^{<3>}$ what's happening there	$k^{<3>}$	$v^{<3>}$
$q^{<4>}$	$k^{<4>}$	$v^{<4>}$
$q^{<5>}$	$k^{<5>}$	$v^{<5>}$

$$q^{<3>} = W^Q \cdot x^{<3>}$$
$$k^{<3>} = W^K \cdot x^{<3>}$$
$$v^{<2>} = W^V x^{<3>}$$

[Vaswani et al. 2017, Attention Is All You Need]

Andrew Ng

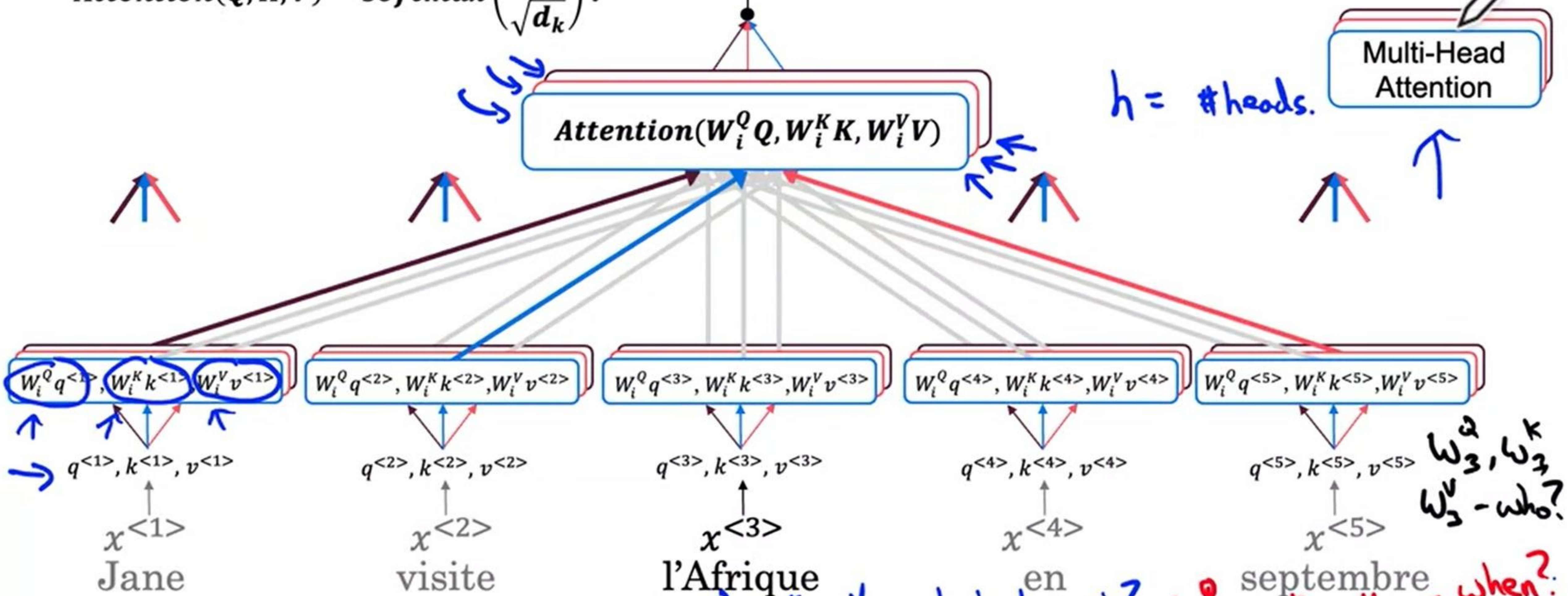
Multi-Head Attention

"head"

$$\boxed{MultiHead(Q,K,V)} = concat(\underline{head_1 head_2 \dots head_h})W_o$$

$$head_i = Attention(W_i^Q Q, W_i^K K, W_i^V V)$$

$$\boldsymbol{Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V}$$



Multi-Head Attention
$Attention(W_i^Q Q, W_i^K K, W_i^V V)$
$W_i^Q q^{<1>}, W_i^K k^{<1>}, W_i^V v^{<1>}$
$W_i^Q q^{<2>}, W_i^K k^{<2>}, W_i^V v^{<2>}$
$W_i^Q q^{<3>}, W_i^K k^{<3>}, W_i^V v^{<3>}$
$W_i^Q q^{<4>}, W_i^K k^{<4>}, W_i^V v^{<4>}$
$W_i^Q q^{<5>}, W_i^K k^{<5>}, W_i^V v^{<5>}$
$q^{<1>}, k^{<1>}, v^{<1>}$
$q^{<2>}, k^{<2>}, v^{<2>}$
$q^{<3>}, k^{<3>}, v^{<3>}$
$q^{<4>}, k^{<4>}, v^{<4>}$
$q^{<5>}, k^{<5>}, v^{<5>}$
$x^{<1>}$ Jane
$x^{<2>}$ visite
$x^{<3>}$ l'Afrique
$x^{<4>}$ en
$x^{<5>}$ septembre

h = #heads.

W_3^Q, W_3^K, W_3^V - who?

W_1^Q, W_1^K, W_1^V - what's happening?

W_2^Q, W_2^K, W_2^V — when?

[Vaswani et al. 2017, Attention Is All You Need]

Andrew Ng

Translation Memory + NMT –> LLM (1) 60%

EN:
The Coordinating Authority receiving the complaint shall assess the complaint and, where appropriate, transmit it to the Coordinating Authority of establishment.

eTranslation:
Koordinačný orgán, ktorému bola sťažnosť doručená, posúdi sťažnosť a v prípade potreby ju postúpi koordinačnému orgánu v mieste usadenia.

Translation Memory (TM):
Koordinačný orgán, ktorý prijal sťažnosť, ju posúdi a podľa potreby postúpi koordinačnému orgánu v krajine usadenia, pričom k nej, ak to považuje za vhodné, pripojí stanovisko.

TM+NMT–>LLM:
Koordinačný orgán, ktorý prijal sťažnosť, ju posúdi a podľa potreby postúpi koordinačnému orgánu v krajine usadenia.

Translation Memory + NMT –> LLM (2) 30%

EN:
A request for confidentiality shall be accompanied by a non-confidential summary of the information supplied or by a statement of the reasons why the information cannot be summarised in a non-confidential manner.

eTranslation:
K žiadosti o zachovanie dôvernosti sa pripojí zhrnutie poskytnutých informácií, ktoré nemá dôverný charakter, alebo vyhlásenie o dôvodoch, prečo informácie nemožno zhrnúť spôsobom, ktorý nemá dôverný charakter.

Translation Memory (TM):
K žiadosti o dôverné zaobchádzanie sa pripojí a musí byť k nim pripojené zhrnutie informácií, ktoré nemá dôvernú povahu, alebo vyhlásenie o dôvodoch, prečo tieto informácie nie je možné zhrnúť.

TM+NMT–>LLM:
K žiadosti o dôverné zaobchádzanie sa pripojí nezverejnené zhrnutie informácií alebo vyhlásenie o dôvodoch, prečo tieto informácie nie je možné zhrnúť v nezverejnenej forme.

Translation Memory + NMT –> LLM (3) 10%

EN:
President of the European Council has a speech to the audience.

eTranslation:
Predseda Európskej rady vystúpi s publikom.

Translation Memory (TM):
Predsedníčka Európskej komisie vystupuje s príhovorom k publiku.

TM+NMT–>LLM:
Predsedník Európskej rady vystupuje s príhovorom k publiku.

Can LLM (like ChatGPT) understand? (1)

A farmer with a wolf, a goat, and a carrot must cross a river by boat. The boat can carry only the farmer and a single item.

1. goat can eat the wolf.
2. wolf can eat a carrot.
3. carrot eats the wolf and the goat

How can they cross the river without anything being eaten?

ChatGPT-4 answer:

Your problem has a twist where the interactions are different from the traditional version of this problem:

1. Goat can eat the Wolf.
2. Wolf can eat the Carrot.
3. Carrot can eat both the Wolf and the Goat.

Can LLM (like ChatGPT) understand? (2)

Here's how the farmer can get all three across the river without anything being eaten:

1. Take the Goat across the river and leave it on the other side (leaving the Wolf and the Carrot on the original side).
2. Return alone to the original side and take the Wolf across the river.
3. Leave the Wolf on the other side, but take the Goat back to the original side.
4. Leave the Goat on the original side and take the Carrot across the river.
5. Leave the Carrot on the other side with the Wolf (since neither can eat each other).
6. Return to the original side one last time to get the Goat.
7. Finally, bring the Goal to the other side.

Now, all three items - the Goat, the Wolf, and the Carrot are on the other side of the river, and none of them have been eaten.

Artificial Neural Networks...

…hallucinations difficult to fix

Zdroje:
[1] Google Research, 2015
[2] Simonite, WIRED, 2018
[3] Katte, Analytics India, 2018
[4] Jun-Yan Zhu et alii, ICCV 2017, arXiv, 2018

NLP research at the DGT (1)

- Daniel Kluvanec:
  Getting the right mix – Approaches to machine translation in the European Commission;
  in Proceedings of the XXth FIT World Congress, Berlin, Germany, 2014
- Patrick Cadwell, Sheila Castilho, Sharon O'Brien, Linda Mitchell:
  Human factors in machine translation and post-editing among institutional translators;
  Translation Spaces 5(2):222-243, 2016
- László Tihanyi, Csaba Oravecz:
  First Experiments and Results in English-Hungarian Neural Machine Translation;
  in Proceedings of the 13th Conference on Hungarian Computational Linguistics (MSZNY 2017), 275-286, Szeged, Hungary, 2017
- Adrien Lardilleux, Yves Lepage:
  CHARCUT: Human-Targeted Character-Based MT Evaluation with Loose Differences;
  In Proceedings of the 14th International Conference on Spoken Language Translation, 146-153, Tokyo, Japan, IWSLT 2017
- Patrick Cadwell, Sharon O’Brien, Carlos S. C. Teixeira:
  Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators;
  Perspectives, Studies in Translation Theory and Practice, Volume 26, 2018 - Issue 3

NLP research at the DGT (2)

- Csaba Oravecz, Katina Bontcheva, Adrien Lardilleux, László Tihanyi, Andreas Eisele:
  eTranslation’s Submissions to the WMT 2019, News Translation Task;
  In Proceedings of the 4th Conference on Machine Translation, 320–326, Florence, Italy, ACL 2019
- Csaba Oravecz, Katina Bontcheva, László Tihanyi, David Kolovratnik, Bhavani Bhaskar, Adrien Lardilleux, Szymon Klocek, Andreas Eisele:
  eTranslation’s Submissions to the WMT 2020, News Translation Task;
  In Proceedings of the 5th Conference on Machine Translation, 254–261, Online, ACL 2020
- Csaba Oravecz, Katina Bontcheva, David Kolovratník, Bhavani Bhaskar, Michael Jellinghaus, Andreas Eisele:
  eTranslation’s Submissions to the WMT 2021, News Translation Task;
  In Proceedings of the 6th Conference on Machine Translation, 72–179, Online, ACL 2021
- Csaba Oravecz, Katina Bontcheva, David Kolovratník, Bogomil Kovachev, Christopher Scott:
  eTranslation’s Submissions to the WMT 2022, General MT Task;
  In Proceedings of the 7th Conference on Machine Translation, pages 346–351, Abu Dhabi, United Arab Emirates (Hybrid), ACL 2022

High Performance Computing (HPC)

- **Knowledge Distillation**
  Csaba Oravecz, Bhavani Bhaskar, Katina Bontcheva, Bogomil Kovachev: Building High Capacity Teacher Models on HPC Infrastructures for the eTranslation Service;
  [ASLING: Translating and the Computer – TC45, Nov. 2023]

- **Large Language Models for Generative Artificial Intelligence in the public sector**

The Project is funded by the European Union
MELUXINA
HIGH PERFORMANCE COMPUTING IN LUXEMBOURG

Contact: [DGT-ETRANSLATION-ADVISORY@ec.europa.eu] (eTranslation) & [EC-AI-NETWORK@ec.europa.eu] (HPC & AI)

Thank you!

© European Union 2023

Unless otherwise noted the reuse of this presentation is authorised under the CC BY 4.0 license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

Sources:
Slide 8: TEF, Tralogy 2013, STOA 2013, FIT 2014, JIAMCATT 2016, TIC 2018, Slovko 2023; Slide 9: CIUTI 2017; Slide 10: University of Oxford 2013, 2023; Slide 11: Gartner; Slide 12: Omniscien Technologies; Slide 13: A Survey of Large Language Models 2023; Slide 14: Google Brain; Slides 15+16: Andrew Ng; Slide 20: Google Research 2015, Simonite WIRED 2018, Katte Analytics India 2018, ICCV 2017; Slide 21: ASLING 2023