From Zero to Production: Baltic-Ukrainian Machine Translation Systems to Aid Refugees

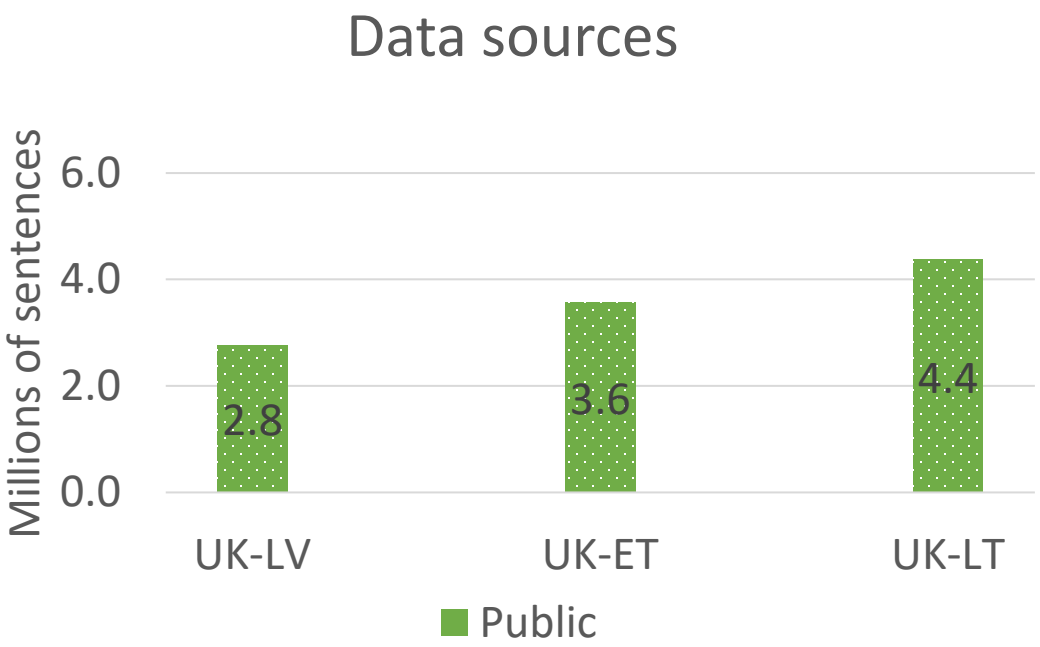
Toms Bergmanis and Mārcis Pinnis

Displaced Ukrainians in Baltics

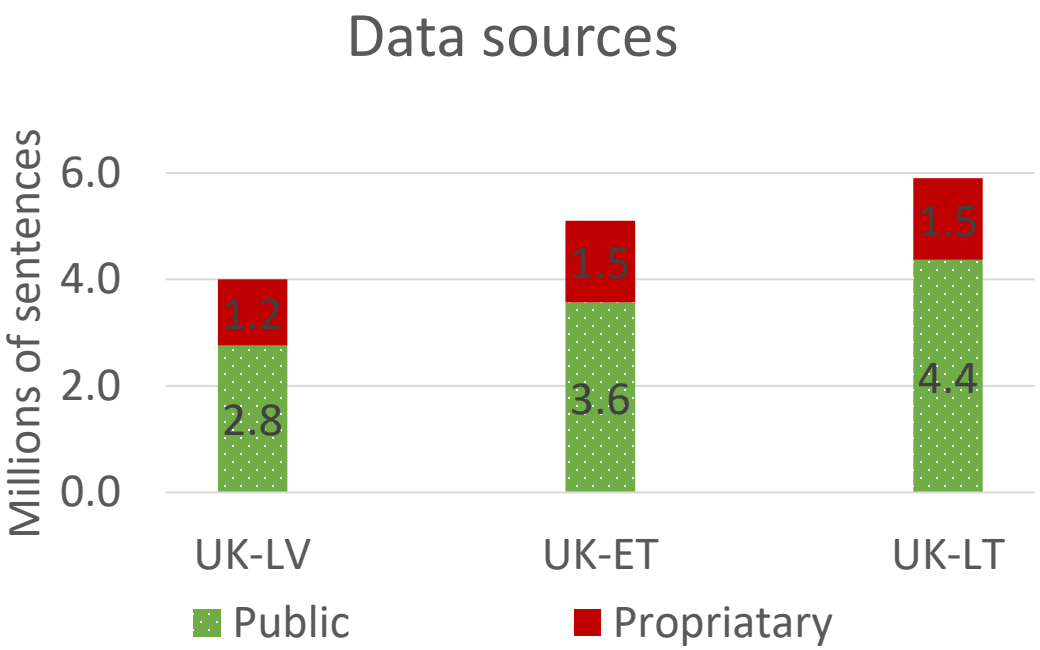
Source https://data.unhcr.org/en/situations/ukraine

One shot attempt at Baltic-Ukrainian MT

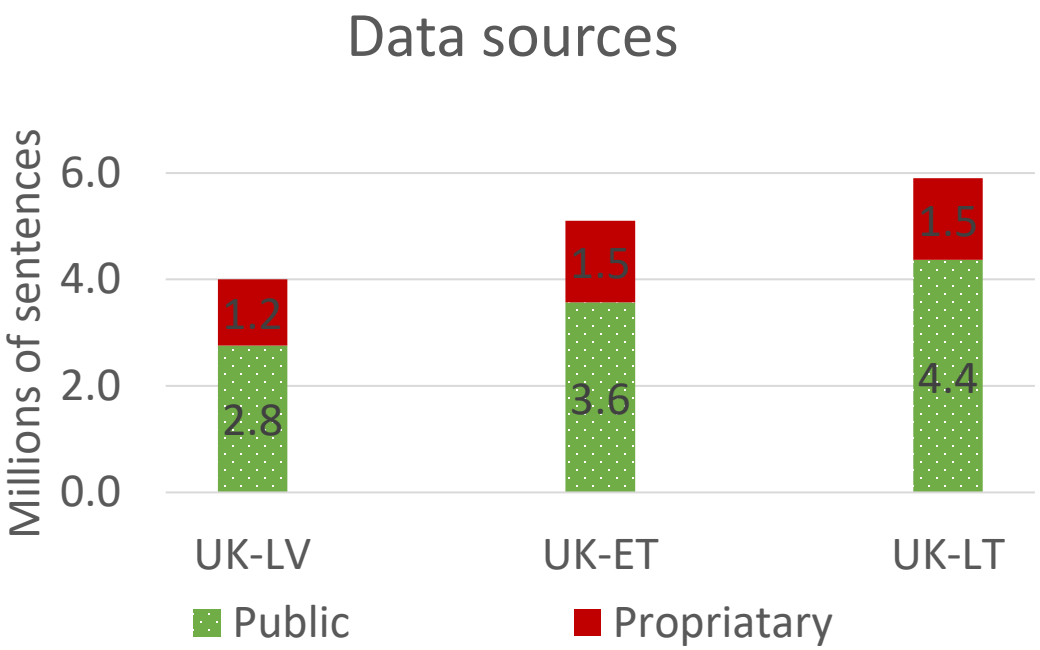
MT System Training Data

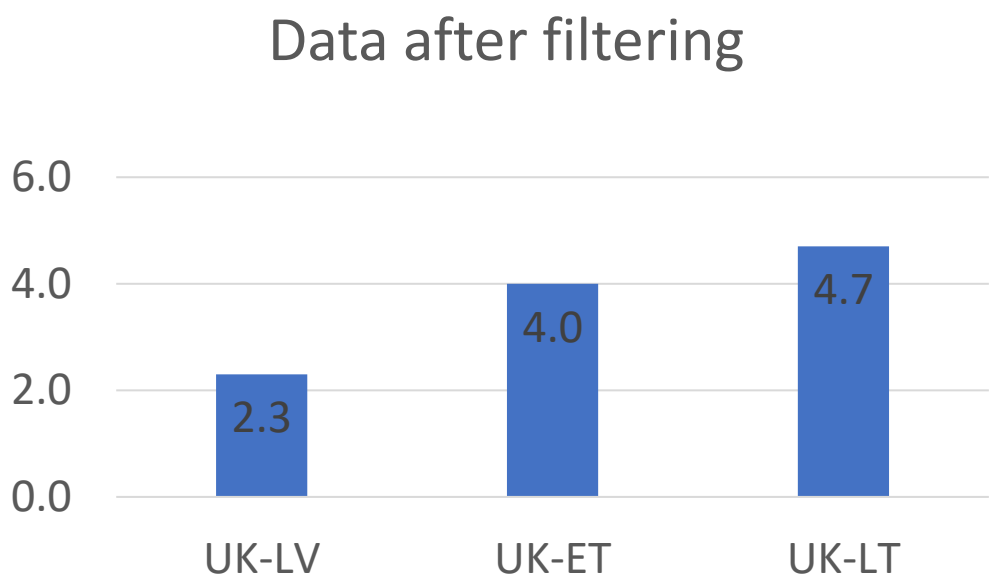


MT System Training Data



MT System Training Data





MT Models

MARIANNMT
FAST NEURAL MACHINE TRANSLATION IN C++

Marian NMT Transformer Base models with support for:

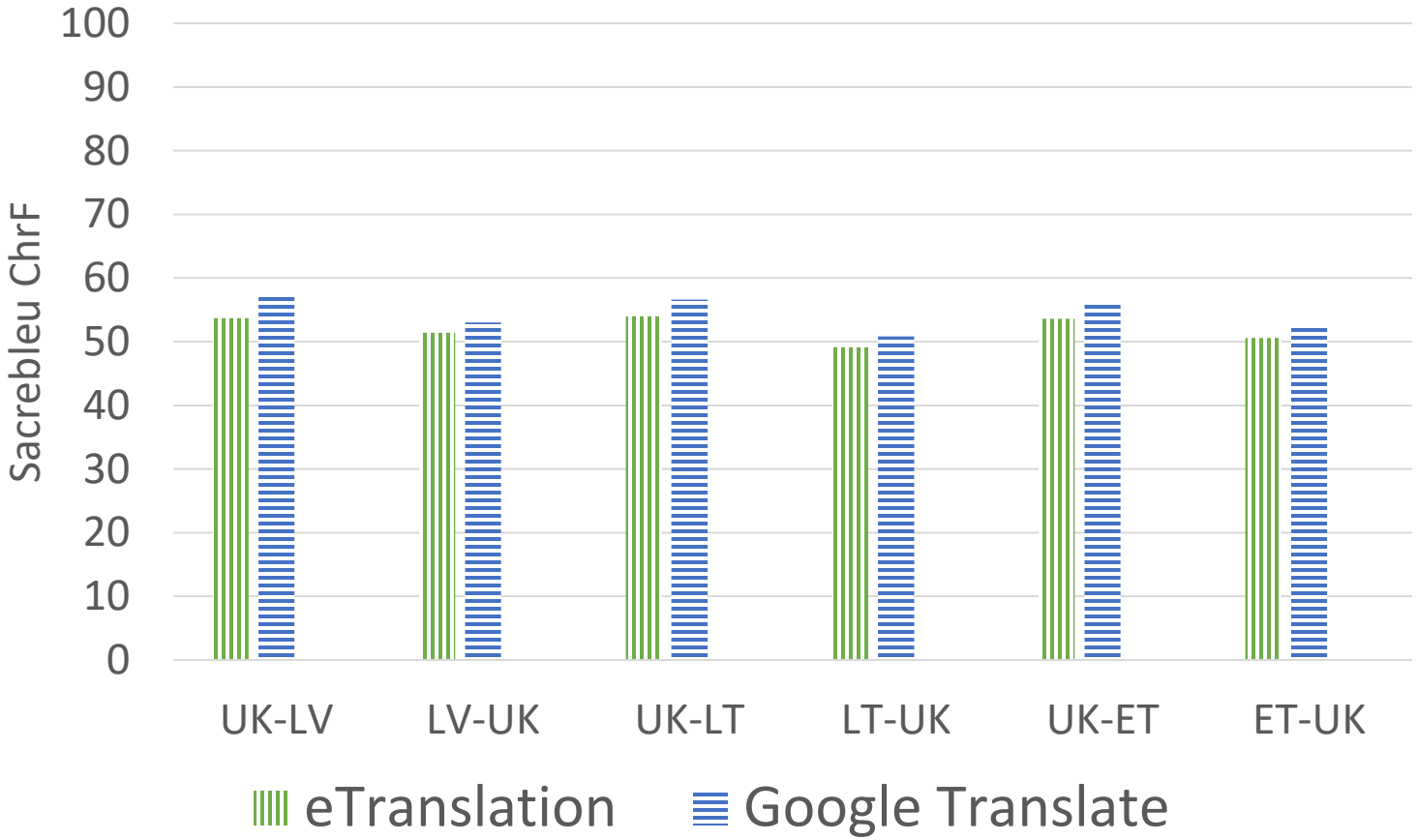
- untranslatable entities [1]
- translation with terminology [2]

[1] Pinnis, Mārcis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. "Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data." In *International conference on text, speech, and dialogue*, pp. 237-245. Springer, Cham, 2017.

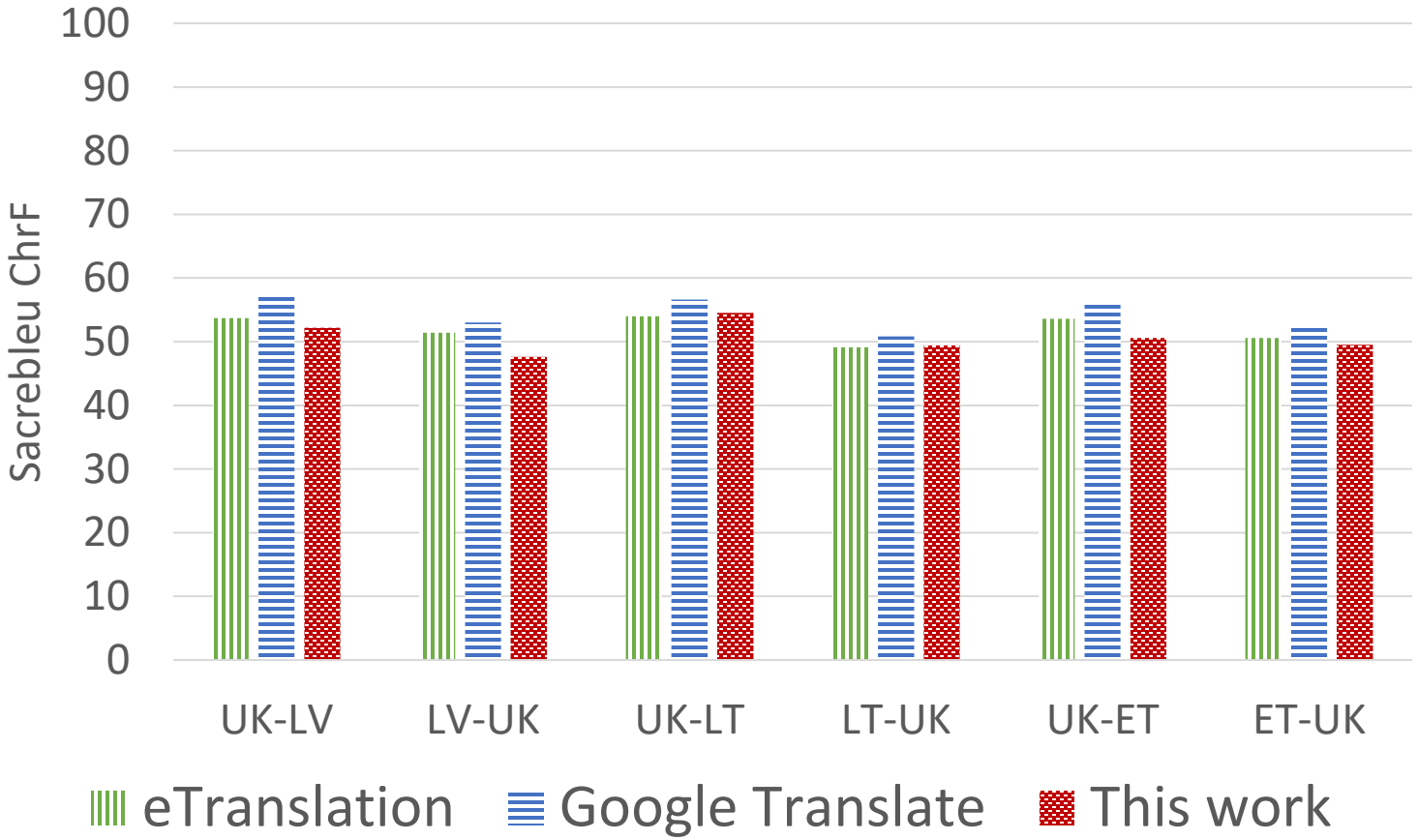
[2] Bergmanis, Toms, and Mārcis Pinnis. "Facilitating Terminology Translation with Target Lemma Annotations." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

System Quality: ChrF metric on FLORES-101

System Quality: ChrF metric on FLORES-101



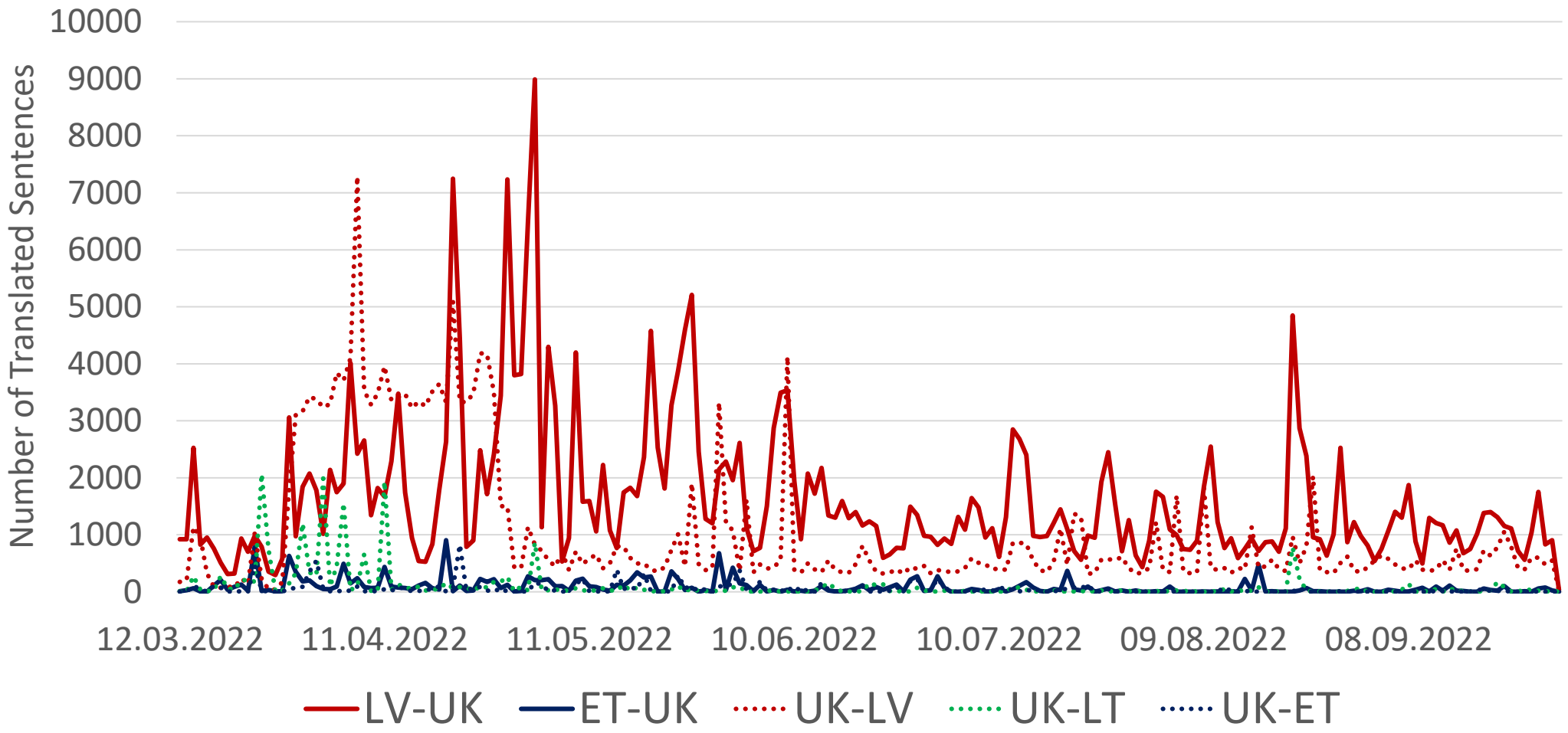
System Quality: ChrF metric on FLORES-101



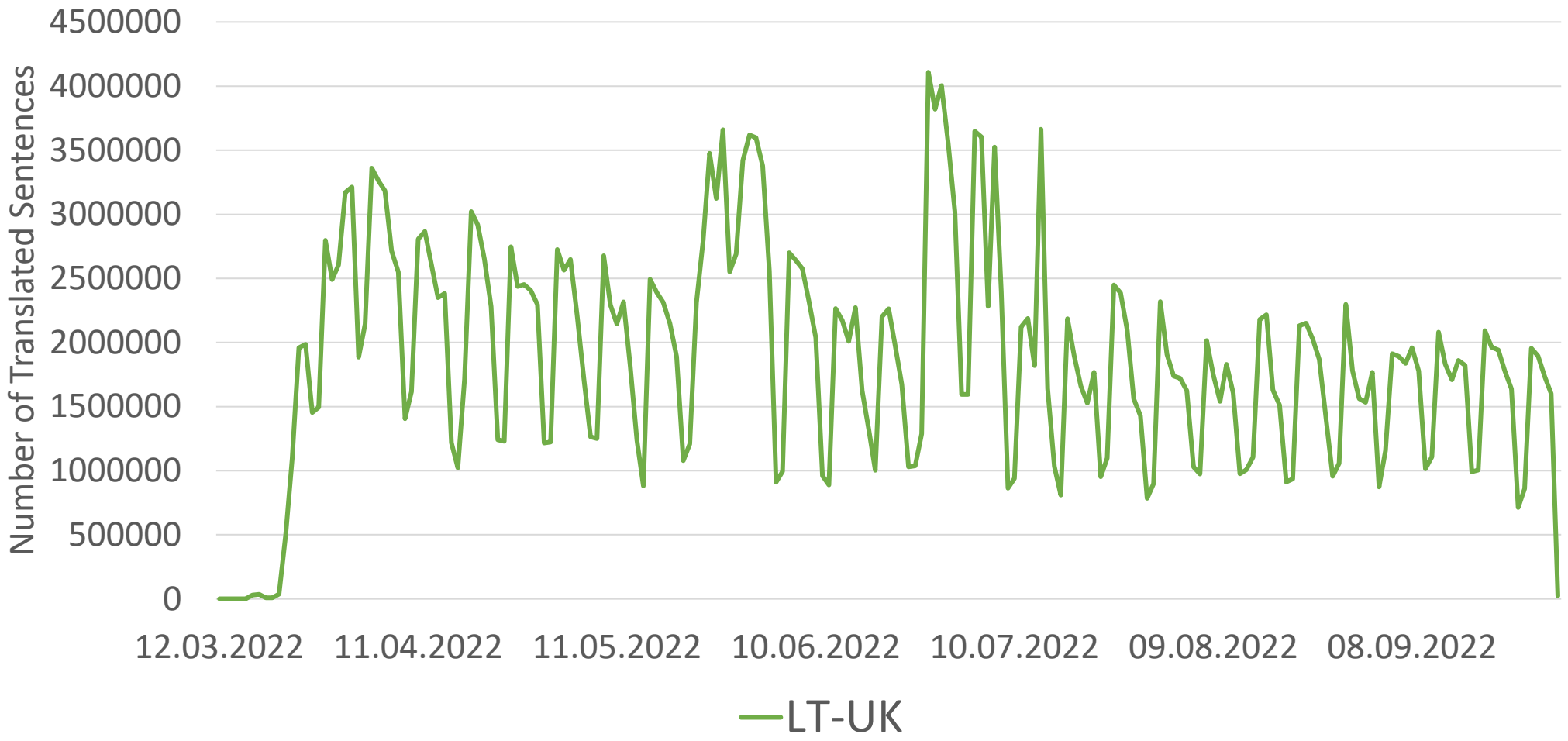
Not too shabby... yet clearly could do with more data

Several months later...

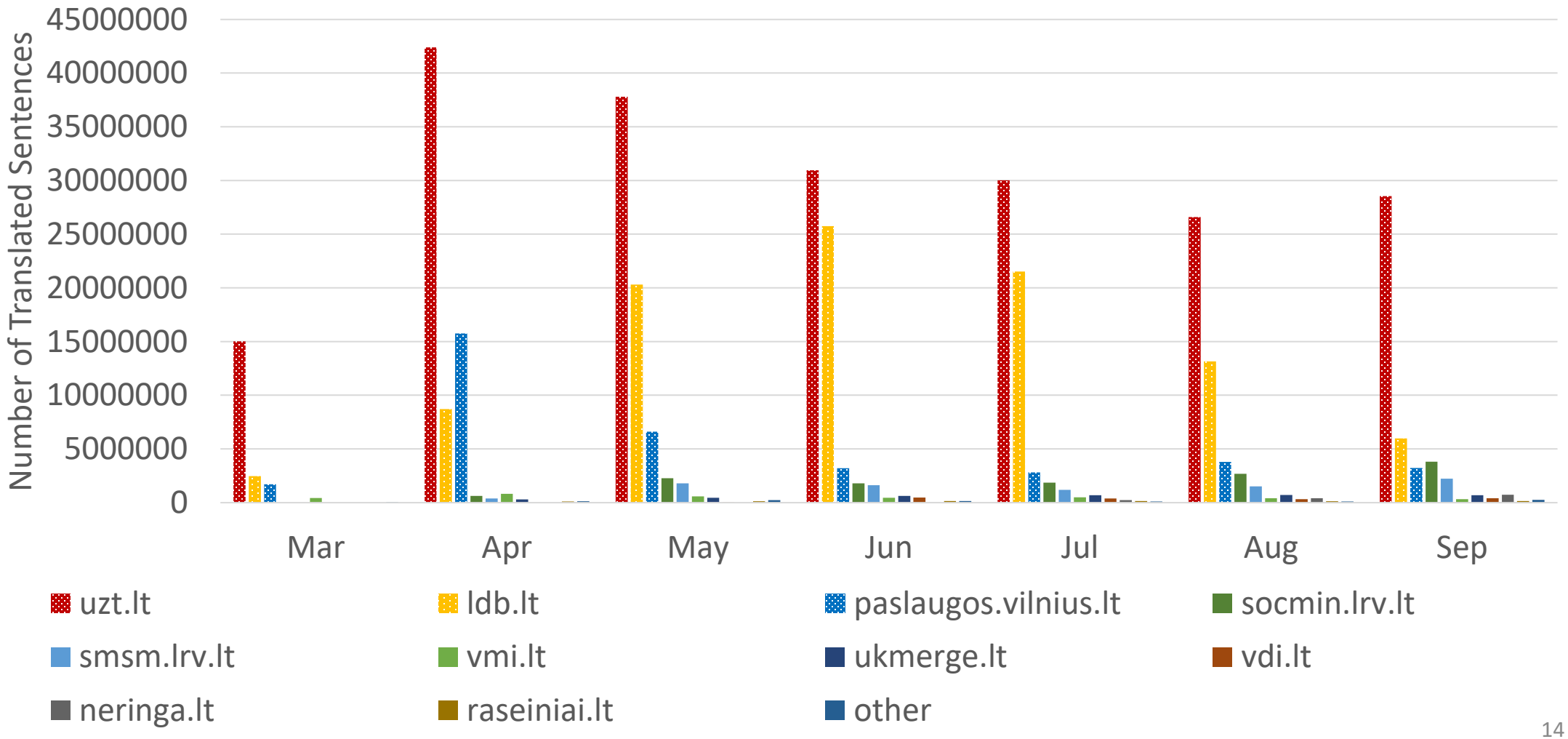
System Usage: all but LT-UK system



System Usage: LT-UK system



LT-UK translation request domain distribution



A step forward

A better LT-UK system

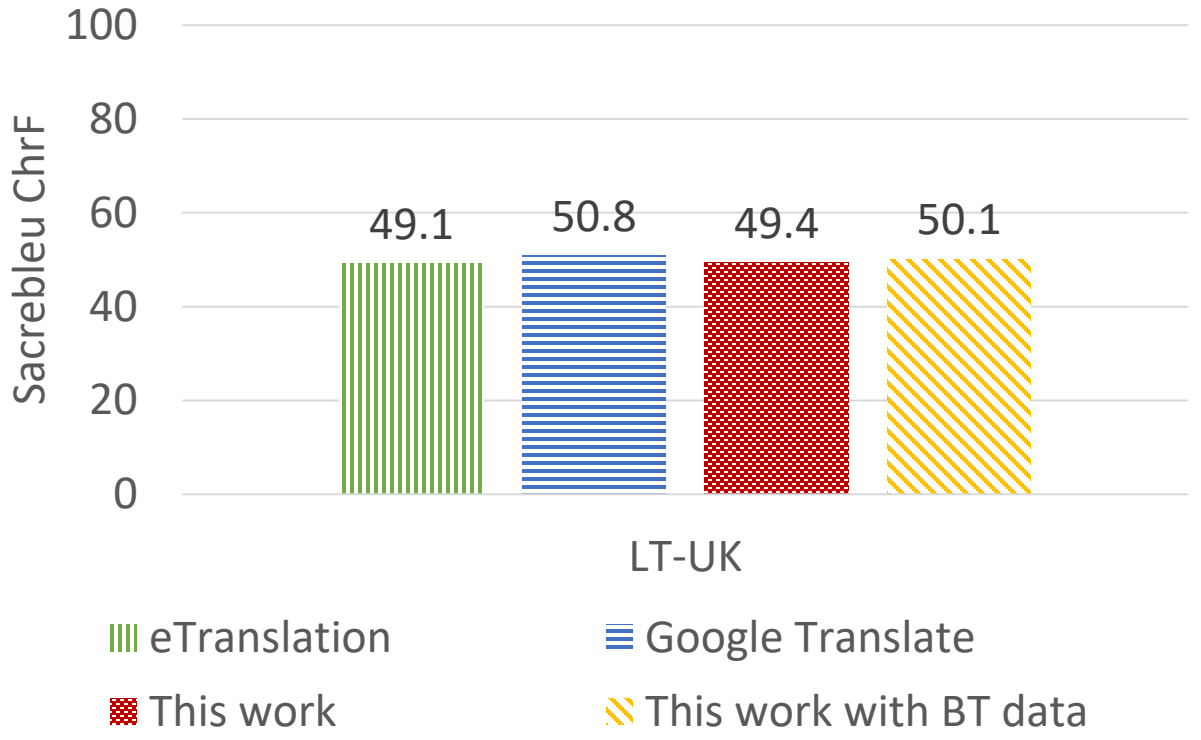
A better LT-UK system

- Back-translated 11.4M Ukrainian monolingual sentences
- Upsampled parallel data in proportion 1:1
- Retrained LT-UK system

Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86-96. 2016.

A better LT-UK system

- Back-translated 11.4M Ukrainian monolingual sentences
- Upsampled parallel data in proportion 1:1
- Retrained LT-UK system



Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86-96. 2016.

Ukrainian Toponyms in Baltic Languages

Examples of Ukrainian Toponyms in Baltic Languages

	English		Latvian		Lithuanian	
	New	**Obsolete**	**New**	**Obsolete**	**New**	**Obsolete**
Київ	Kyiv	Kiev	Kijiva	Kijeva	Kyjivas	Kyjevas
Харків	Kharkiv	Kharkov	Harkiva	Harkova	Charkivas	Charkovas
Дніпро	Dnipro	Dnipropetrovsk	Dnipro	Dņepro	Dnipras	Dniepras
Львів	Lviv	Lvov	Ļviva	Ļvova	Lvivas	Lvovas
Чернігів	Chernihiv	Chernigov	Černihiva	Čerņigova	Černyhivas	Černyhovas

Translation with terminology

Target Lemma Annotations for terminology integration:

Src.: Car has a faulty engine or transmission [..]

Trg.: Automašīnai ir atteice dzinejā vai transmisijas [..]

With TLA: Car|w has|w a|w faulty|w engine|s dzinejējs|t or|w transmission|w

Bergmanis, Toms, and Mārcis Pinnis. "Facilitating Terminology Translation with Target Lemma Annotations." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Translation with toponyms as terminology

Src:	**Київ, Харків, Одеса, Дніпро і Донецьк мають пять найбільших українських міст.**
UK-LV baseline:	Kijeva, Harkova, Odesa, Dņepra un Doņecka ir piecas lielākās Ukrainas pilsētas.
UK-LV with terminology:	Kijiva, Harkiva, Odesa, Dnipro un Donecka ir piecas lielākās Ukrainas pilsētas.

Translation with toponyms as terminology

Src:	**Київ, Харків, Одеса, Дніпро і Донецьк мають пять найбільших українських міст.**
UK-LV baseline:	Kijeva, Harkova, Odesa, Dņepra un Doņecka ir piecas lielākās Ukrainas pilsētas.
UK-LV with terminology:	Kijiva, Harkiva, Odesa, Dnipro un Donecka ir piecas lielākās Ukrainas pilsētas.
Src:	**Ми доставляємо товари в Запоріжжя, Львів, Чернігів та**
UK-LT baseline:	Mes pristatome prekes į Zaporožę, Lvovą, Černigą ir Ternopilį.
UK-LT with terminology:	Prekes pristatome į Zaporižę, Lvovą, Černyhivą ir Ternopilį.

Conclusions

We presented:

- a set of Baltic-Ukrainian MT systems with results comparable to other public MT service providers
- Evidence of at least two of the systems being used by
  - professional translators
  - end-users accessing social services
- two further improvements for
  - the most used system by using backtranslation
  - toponym translation

Thank you!

The research has been supported by the European Regional Development Fund within the research project “AI Assistant for Multilingual Meeting Management” No. 1.1.1.1/ 19/A/082.

TILDE

NATIONAL DEVELOPMENT PLAN 2020

EUROPEAN UNION
European Regional Development Fund

INVESTING IN YOUR FUTURE