moz://a

Training efficient and open source ML models for private translation in Firefox

And how you can help.

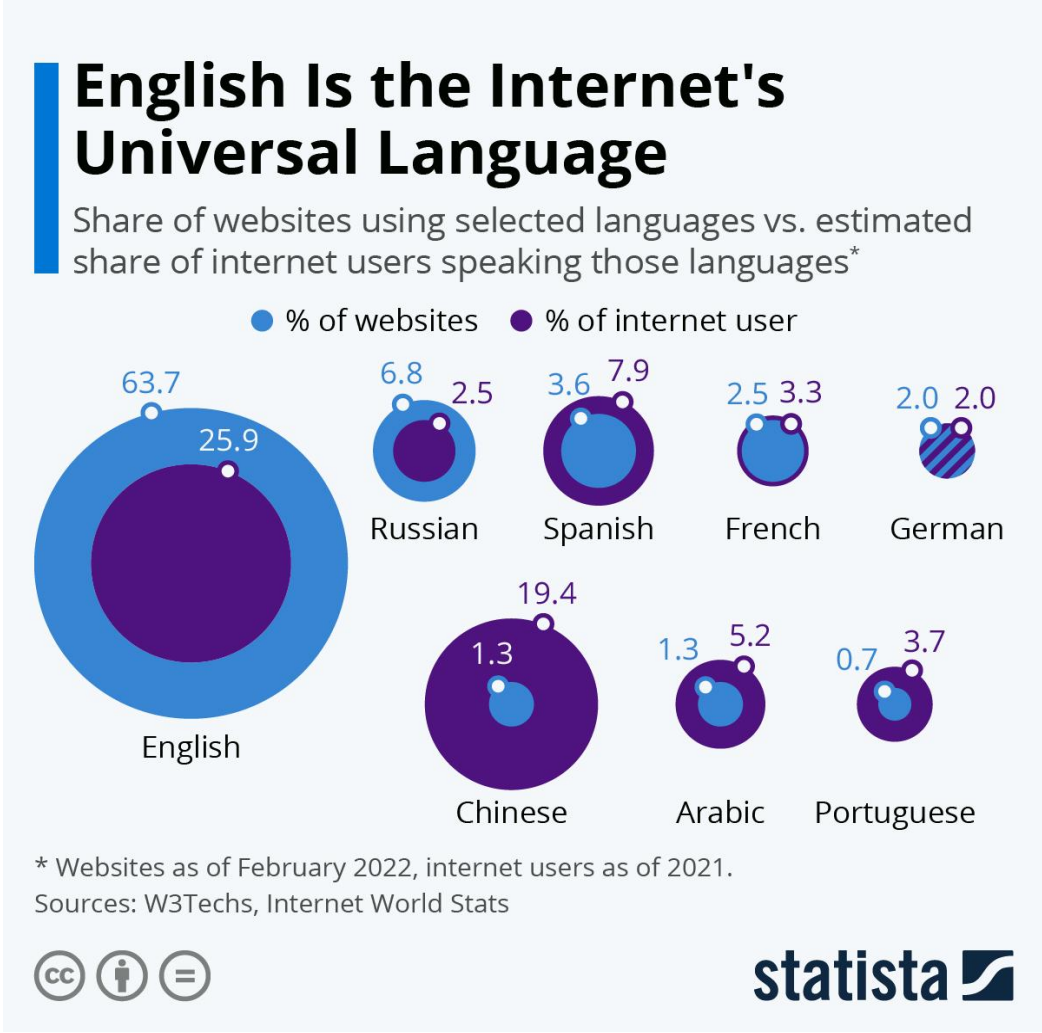
Marco Castelluccio <marco@mozilla.com>

FOSDEM 2024

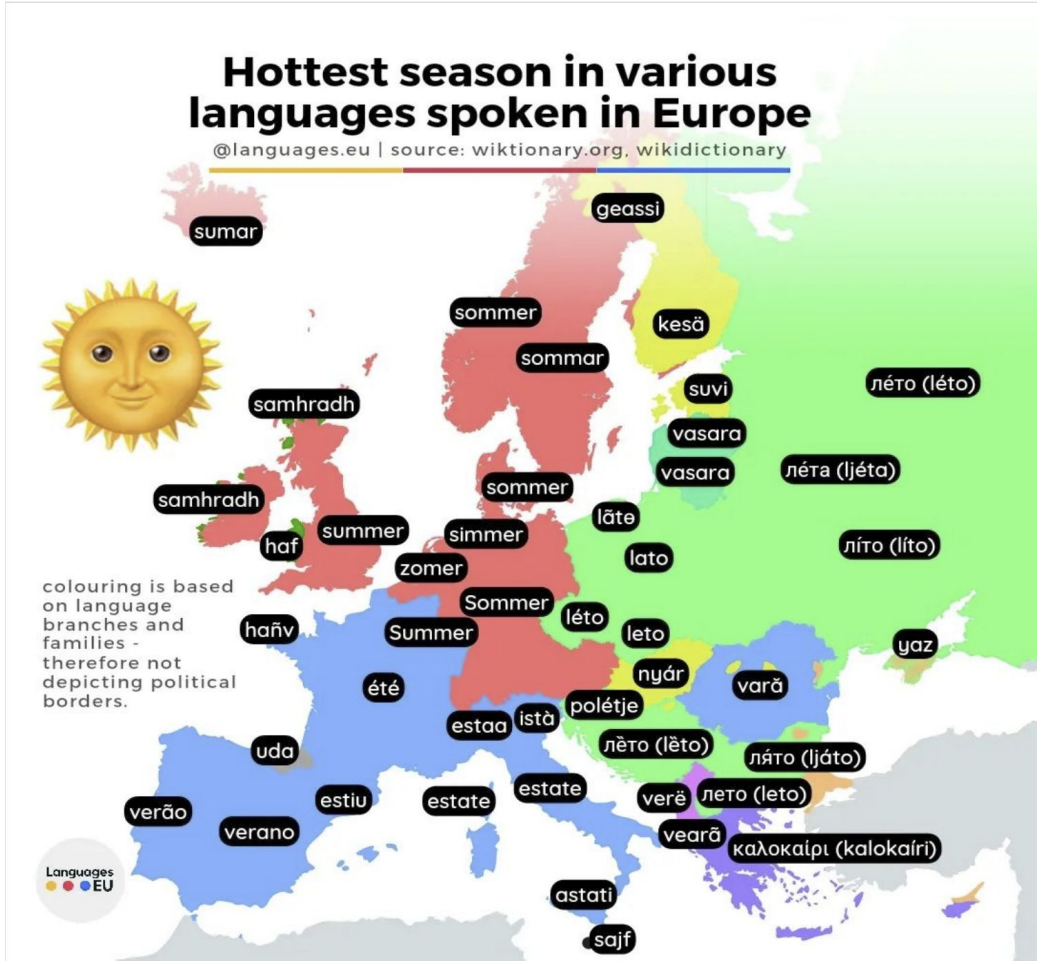
BACKGROUND

- **~64%** of the World Wide Web is in English

- Only **~26%** of users are native English speakers



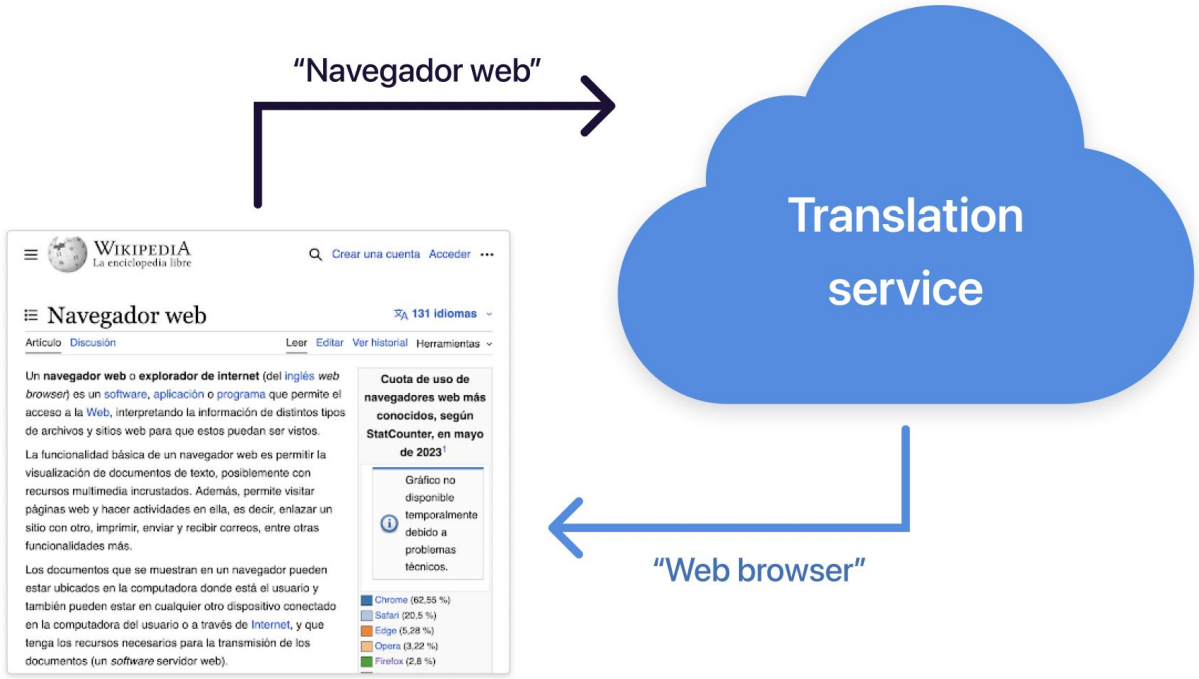
24 official languages in Europe 🇪🇺



93 languages in EU 🇪🇺 spoken by more than 100k people

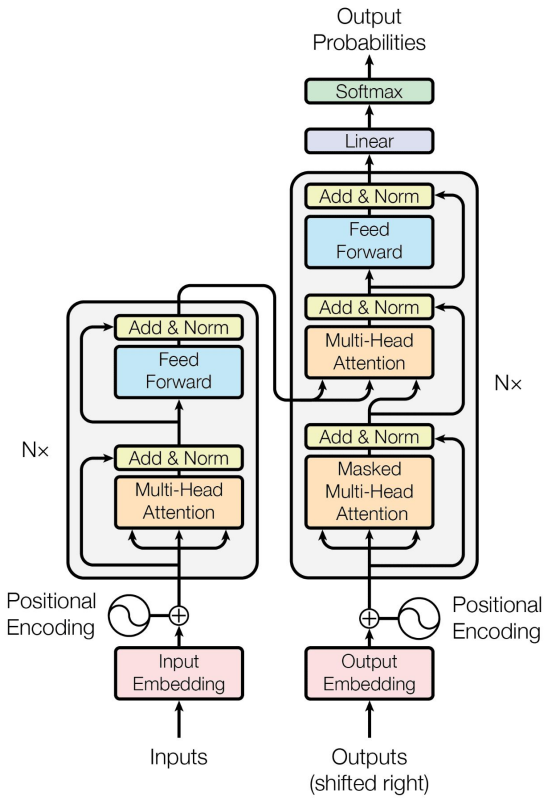
Why privacy-friendly translations

- Current cloud-based services do not offer the **privacy guarantees** that we like to offer in Firefox



Privacy-friendly translations

- **European Union 🇪🇺 funded project** to investigate client-side private translation capabilities



Languages

Released

- Bulgarian, Dutch, English, French, German, Italian, Polish, Portuguese, Spanish

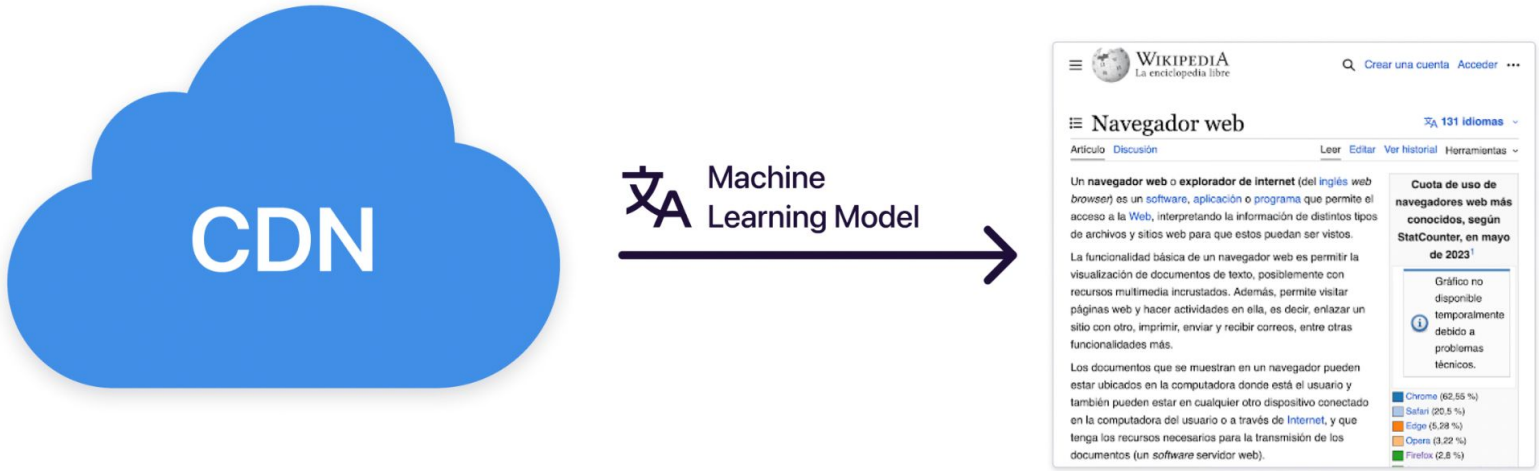
In development

- Catalan, Czech, Estonian, Finnish, Greek, Hungarian, Icelandic, Lithuanian, Maltese, Norwegian Bokmål, Norwegian Nynorsk, Persian (Farsi), Russian, Slovenian, Turkish, Ukrainian

INFERENCE

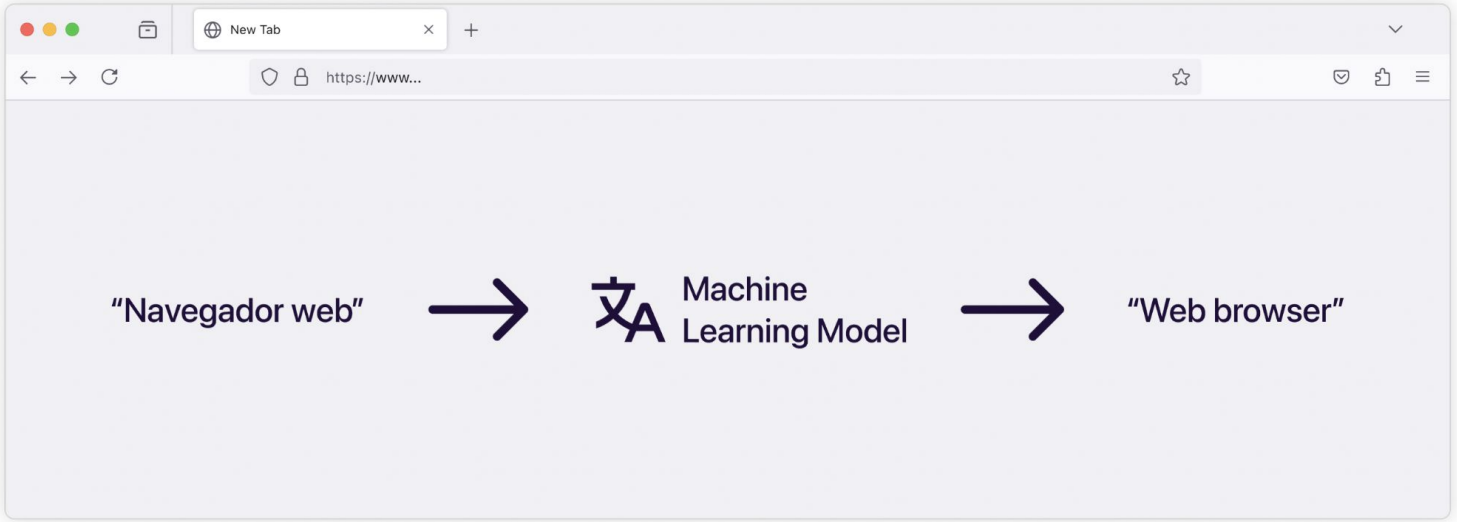
Model download

- Firefox will download and update the model from our servers - just for the languages that you need



Run models locally

- Use **machine learning on the client side** to translate locally
- All the data remains on the system



WASM engine with SIMD optimizations

- NMT engine compiled to **WebAssembly**

- Using **SIMD instructions** via gemmology
  - 10x perf improvement

TRAINING

Open Datasets

An overview of the OPUS collection

1,210 CORPORA

45,945,946,108 TOTAL SENTENCE PAIRS

744 LANGUAGES AVAILABLE

THIS MAP DISPLAYS **10** CORPORA , WHICH MAKE UP A TOTAL **93.40%** OF THE ENTIRE **OPUS** COLLECTION

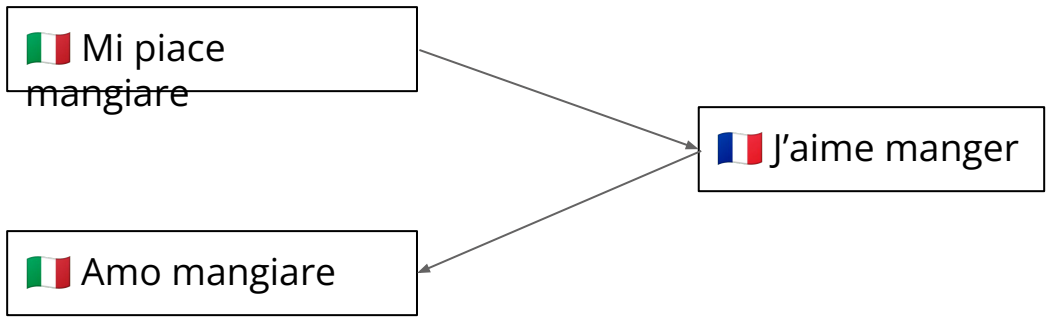
Corpus	Sentences	% of OPUS
NLLB	13B	28.31
CCMatrix	11B	23.64
OpenSubtitles	8.5B	18.53
MultiCCAligned	2.2B	4.87840
ParaCrawl	1.5B	3.26229
DGT	1.1B	2.37845
XLEnt	883M	1.92148
MultiParaCrawl	789M	1.71653
LinguaTools-WikiTitles	487M	1.06082
CCAligned	439M	0.95442

https://opus.nlpl.eu/

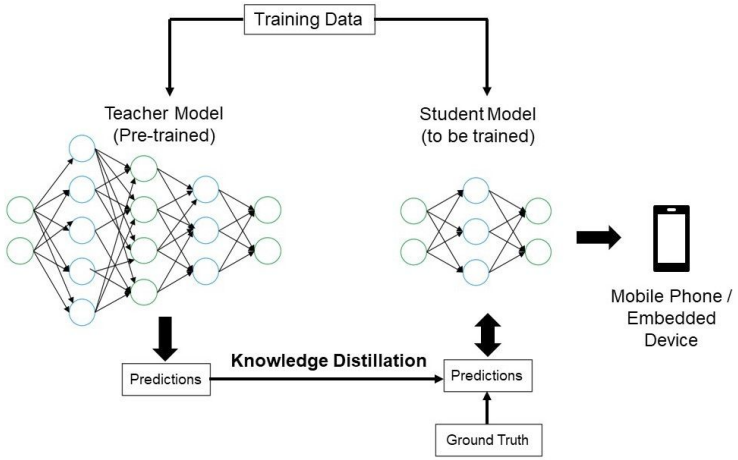
Data cleaning and augmentation

- Rule-based **data cleaning** + ML-based detection of bad sentence pairs (bicleaner-ai)

- **Back-translation** for data augmentation from monolingual data
  - Translate target to source
  - Add pair “noisy translated source” -> “good target”



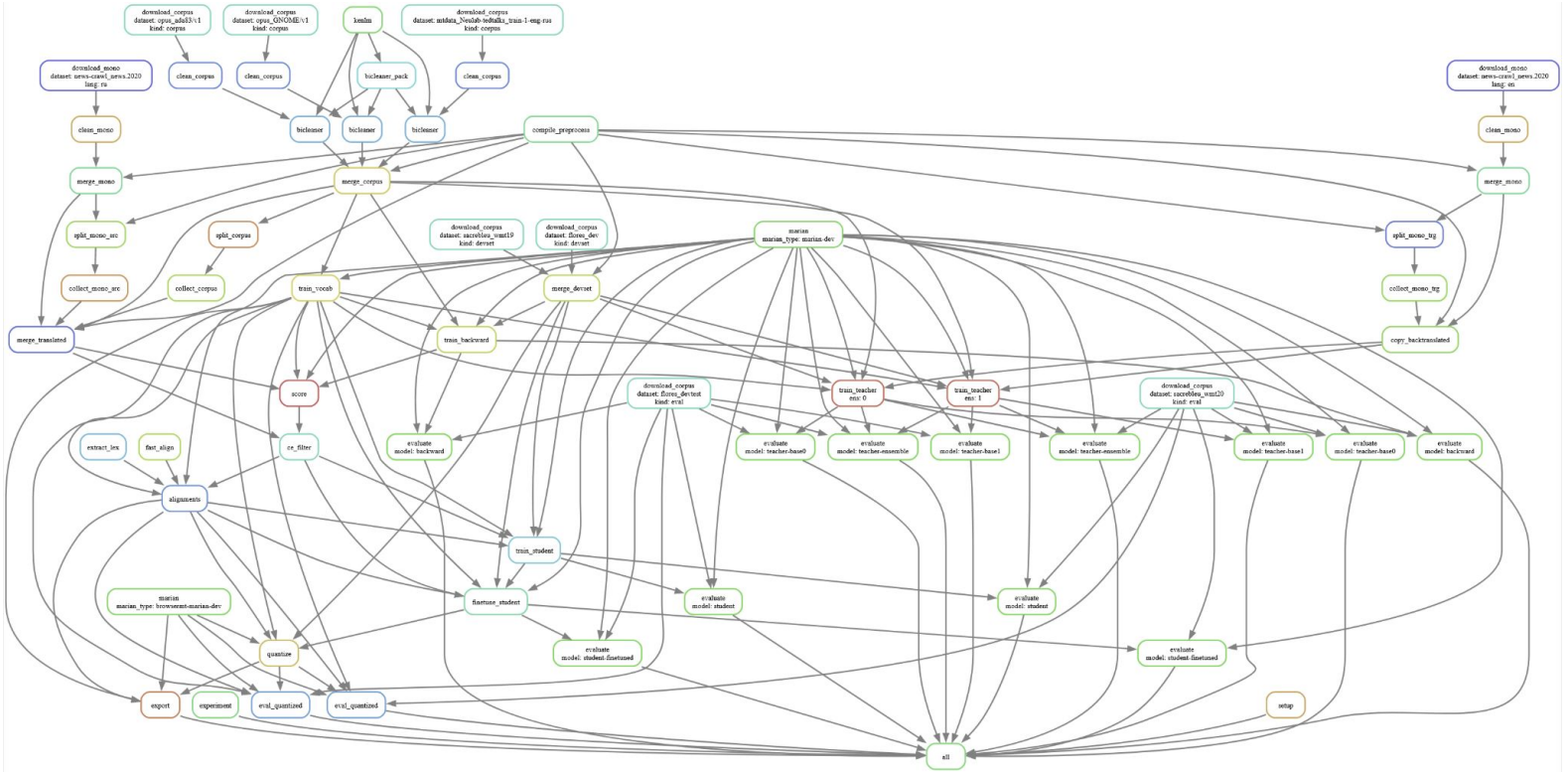
Model training and compression



- **Training** of a large model
- Compression using **knowledge distillation** into a smaller one
- **Quantization** (float32 -> int8) for further compression and perf improvement

Model	Size	Total number of parameters	Dataset decoding time on 1 CPU core	Quality, BLEU
Teacher	798Mb	192.75M	631s	52.5
Student quantized	17Mb	15.7M	17.9s	50.7

Pipeline

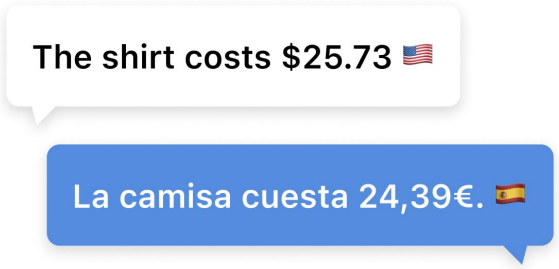


https://mozilla.github.io/firefox-translations-training/pipeline-steps.html

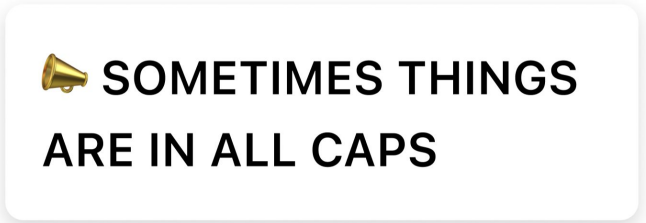
NEXT STEPS

Improving quality

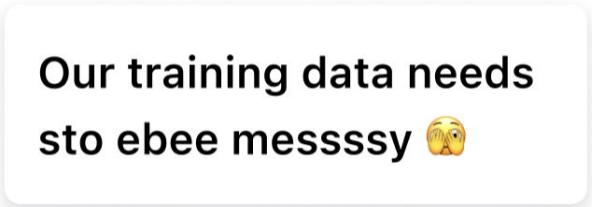
Fix numbers



Fix capitalization



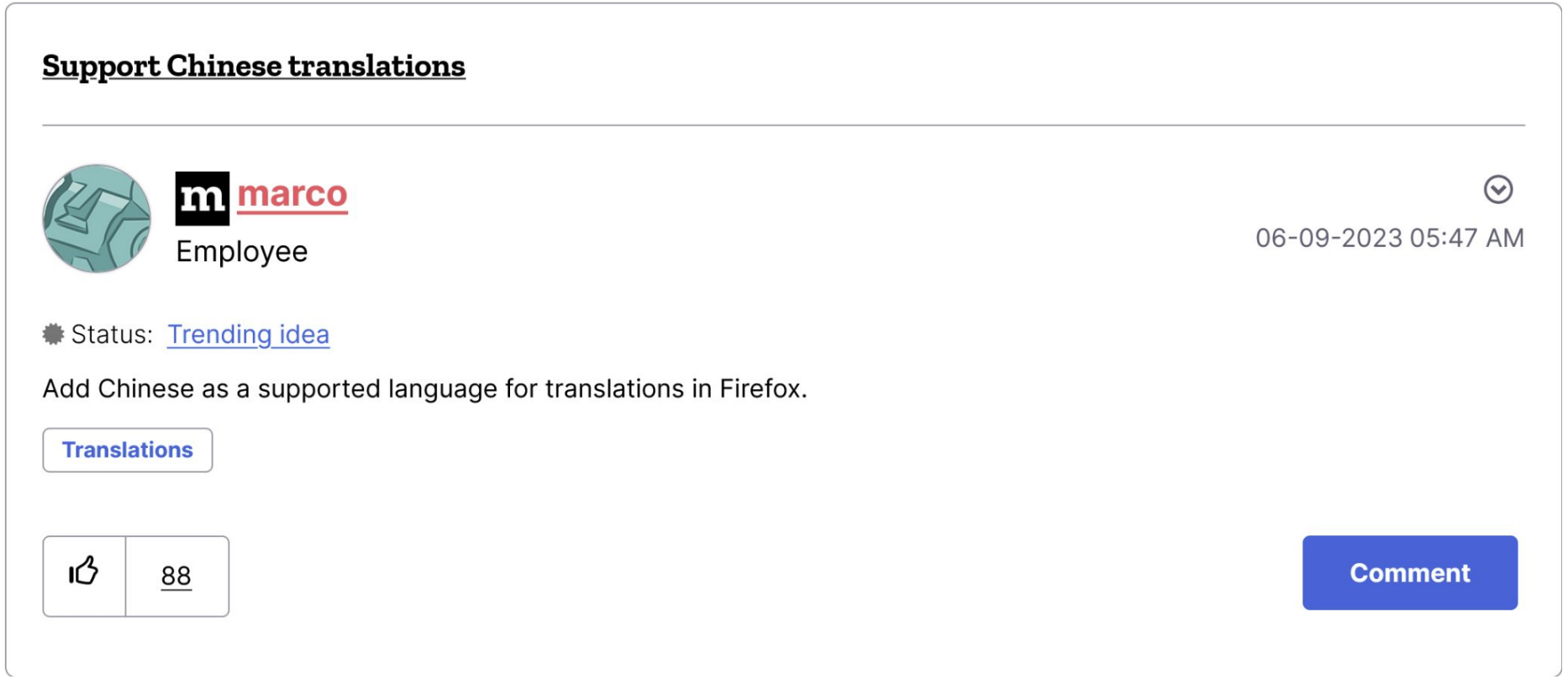
The web is messy!



Experiment	Model	Test Dataset	Augmentation of test dataset	BLEU
	teacher	sacrebleu_wmt19	all upper case	4
	teacher	sacrebleu_wmt19	all title case	6.8
No OpusTrainer augmentation	teacher	sacrebleu_wmt19	no	31.5
	teacher	sacrebleu_wmt19	all title case	31.8
Upper case and title case augmentation with OpusTrainer	teacher	sacrebleu_wmt19	all upper case	29
	teacher	sacrebleu_wmt19	no	31.6
	student	sacrebleu_wmt19	no	29.9
	student	sacrebleu_wmt19	all upper case	28.9
	student	sacrebleu_wmt19	all title case	29.8

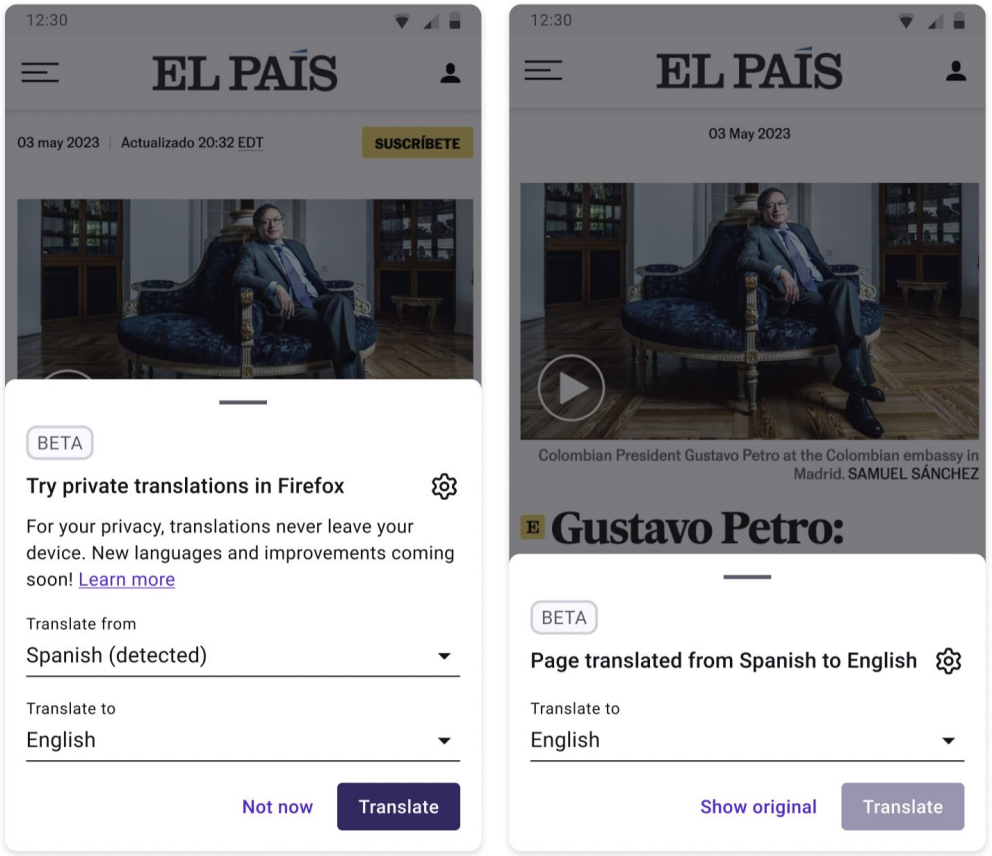
https://github.com/mozilla/firefox-translations-training/issues/216

Supporting more languages



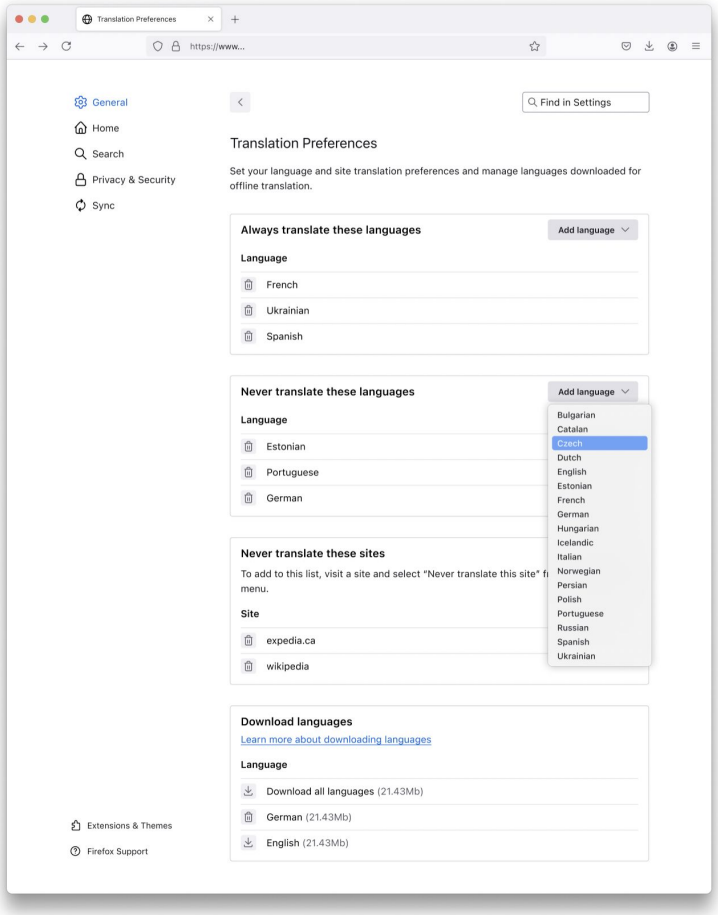
https://github.com/mozilla/firefox-translations-training/issues/369

Android



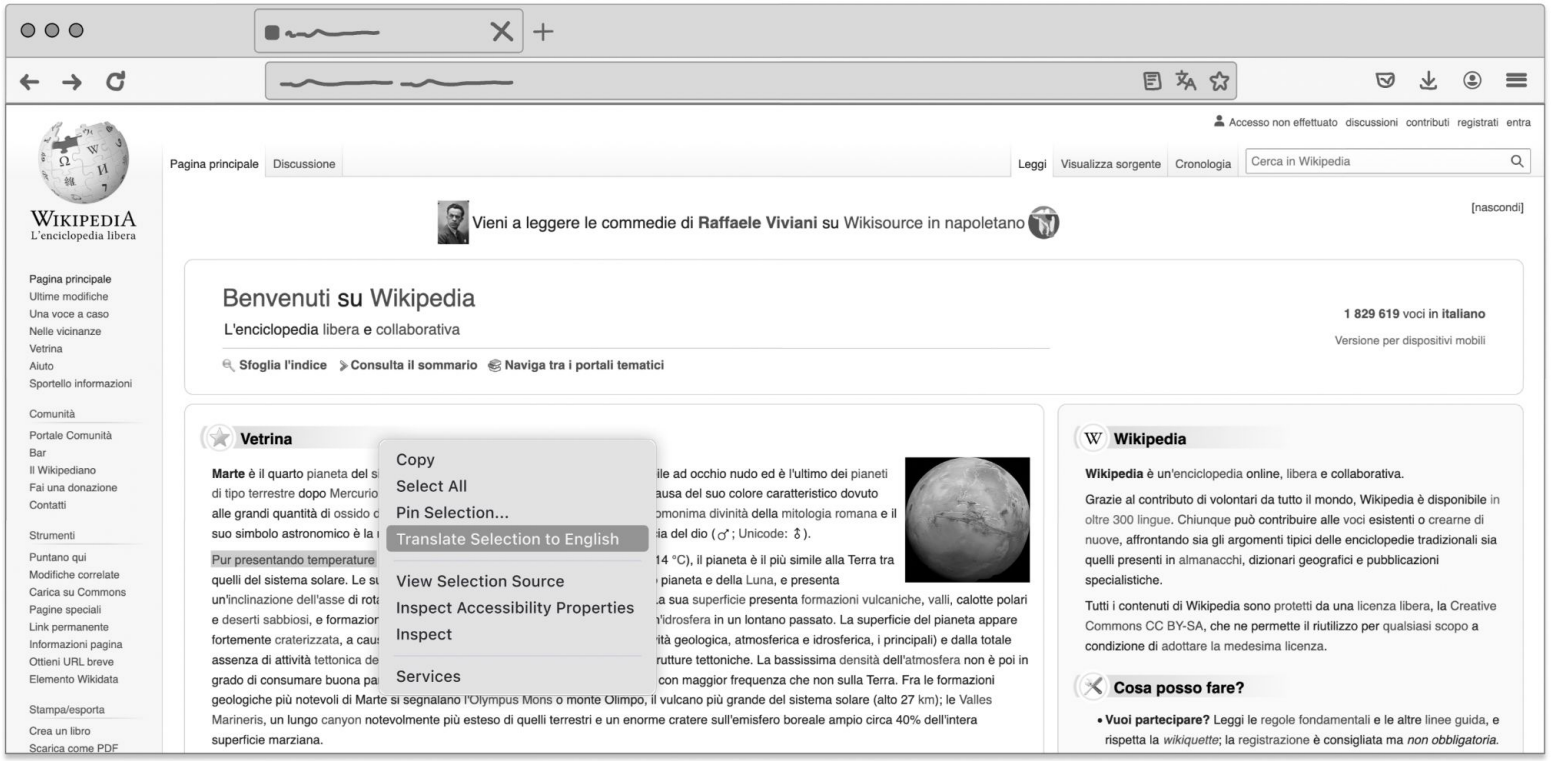
https://bugzilla.mozilla.org/show_bug.cgi?id=1820240

Settings improvements



https://bugzilla.mozilla.org/show_bug.cgi?id=1869015

Selection translation



https://bugzilla.mozilla.org/show_bug.cgi?id=1855907

Help us!

- **Know any dataset** that we could use?
- Want to **contribute code** for a Firefox feature?
- Interested in **adding support for your language**?


grazie 🇮🇹
obrigado 🇵🇹
Благодаря 🇧🇬
dziękuję 🇵🇱
thank you 🇺🇸
gracias 🇪🇸
merci 🇫🇷
danke 🇩🇪
bedankt 🇳🇱