

ADDRESSING NMT CHALLENGES IN TRAINING WITH LIMITED DATA

RIHARDS KRIŠLAUKS
rihards.krislauks@tilde.lv

AI BASED LANGUAGE TECHNOLOGY



NEURAL MACHINE TRANSLATION



NATURAL LANGUAGE PROCESSING



SPEECH PROCESSING



VIRTUAL ASSISTANTS

TILDE

TILDE SUCCESS FACTORS

- RESEARCH & DEVELOPMENT & **DATA CURRATION**
- FAST INNOVATION CYCLE **FROM CUTTING-EDGE RESEARCH TO SCALABLE APPLICATIONS**
- EFFICIENT PAN-EUROPEAN **COLLABORATION**



TILDE MT

Wide range of services

- Enterprise or cloud-based services
- Secure and scalable solution
- General and customer-tailored system development

Rich functionality for state-of-the-art system development

- Data storage and cleaning
- Data pre-processing/post-processing
- SMT and NMT system support for training and decoding
- Document, Web site and text translation
- Translation API for third party integration
- Integrated with the Tilde Terminology platform
- and more...

Popular CAT tool plugins

- Trados SDL Studio
- memoQ
- GlobalSight
- Memsource
- MateCat

CHALLENGES DUE TO LIMITED DATA.



MT requires lots of data

- Good to have 1-4M parallel sentences. **More is** even **better**.
- SMT is as data-hungry as NMT.
- However, SMT has a lot of options to deal with it:
  - Using language models trained on monolingual data;
  - Domain adaptation – using out-of-domain data;
  - Term glossaries;
  - SMT is better at remembering rare events.



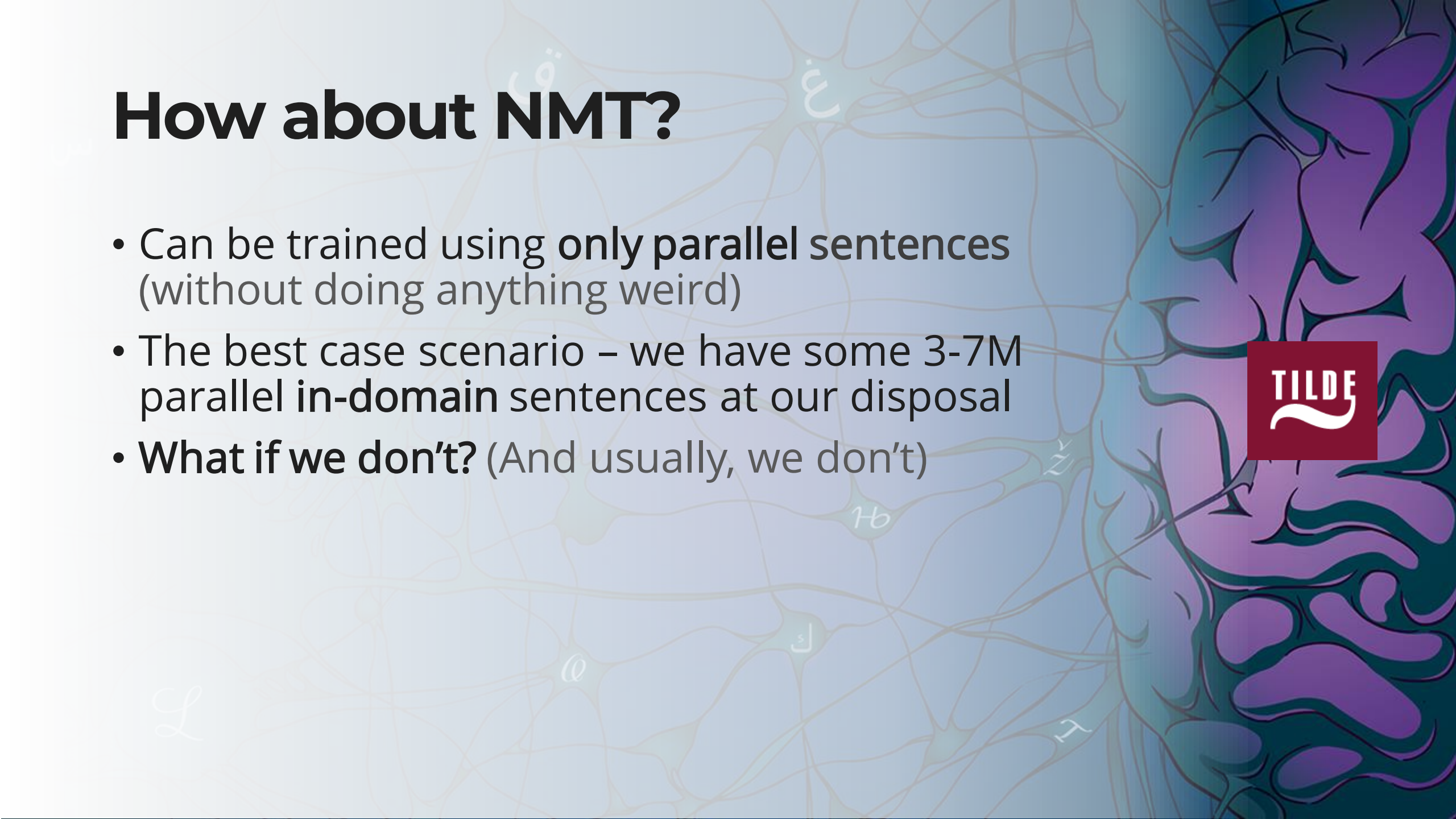
A typical SMT training setup

- 50-100k in-domain sentence pairs
- 3-7M out-of-domain sentence pairs
- 5-20M monolingual sentences (target language)
- 1-2k in-domain sentence pairs for evaluation
- A few hundred terms

SMT involves a bunch of different components trained separately on each of the above types of data

How about NMT?

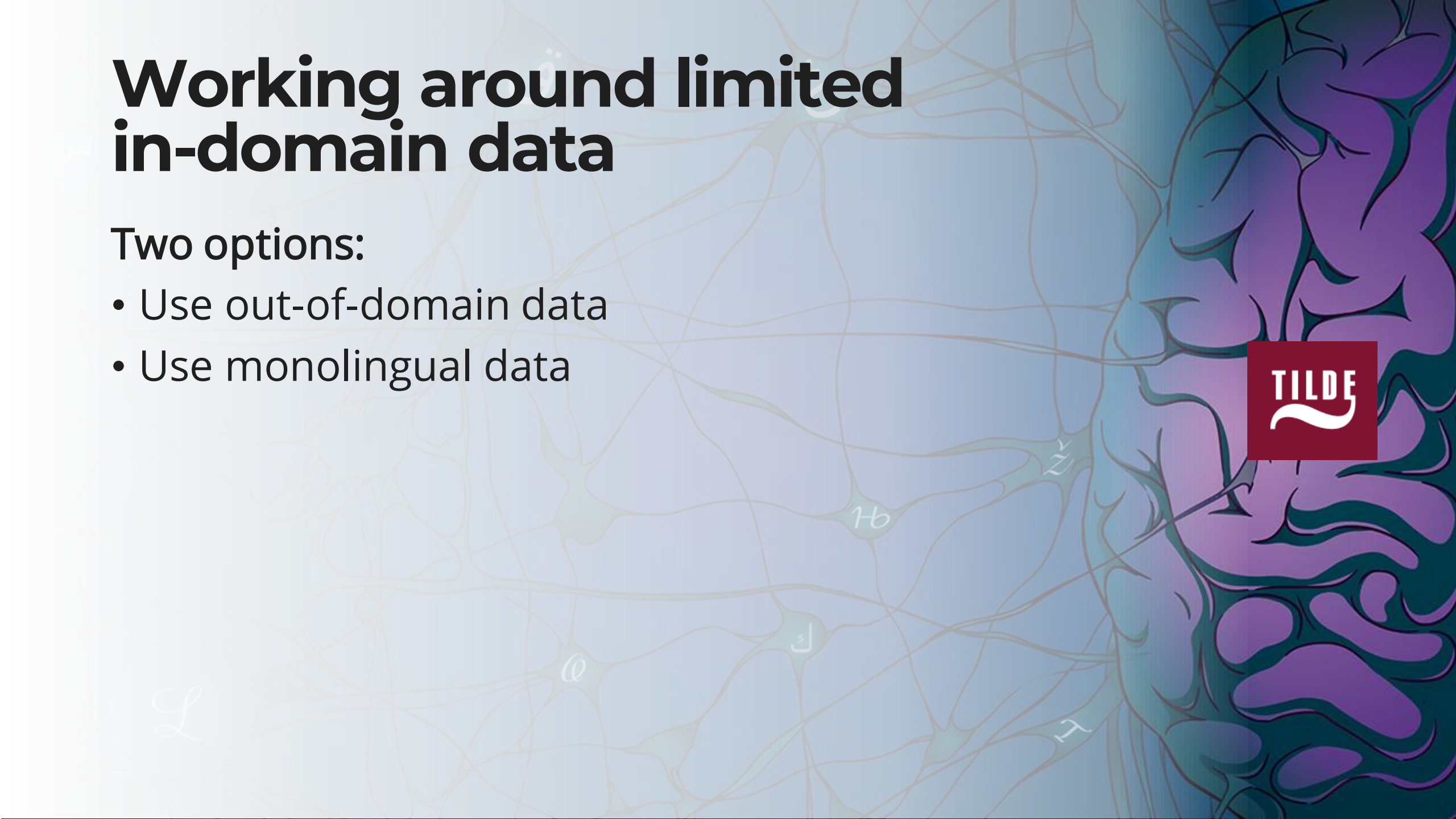
- Can be trained using **only parallel sentences** (without doing anything weird)
- The best case scenario – we have some 3-7M parallel **in-domain** sentences at our disposal
- **What if we don't?** (And usually, we don't)



Working around limited in-domain data

Two options:

- Use out-of-domain data
- Use monolingual data



Using out-of-domain data

1) Mix out-of-domain data with in-domain data during training
   - Lots of options on how to do the “mixing”
2) Pre-train the NMT system on out-of-domain data and tune it on in-domain data afterwards

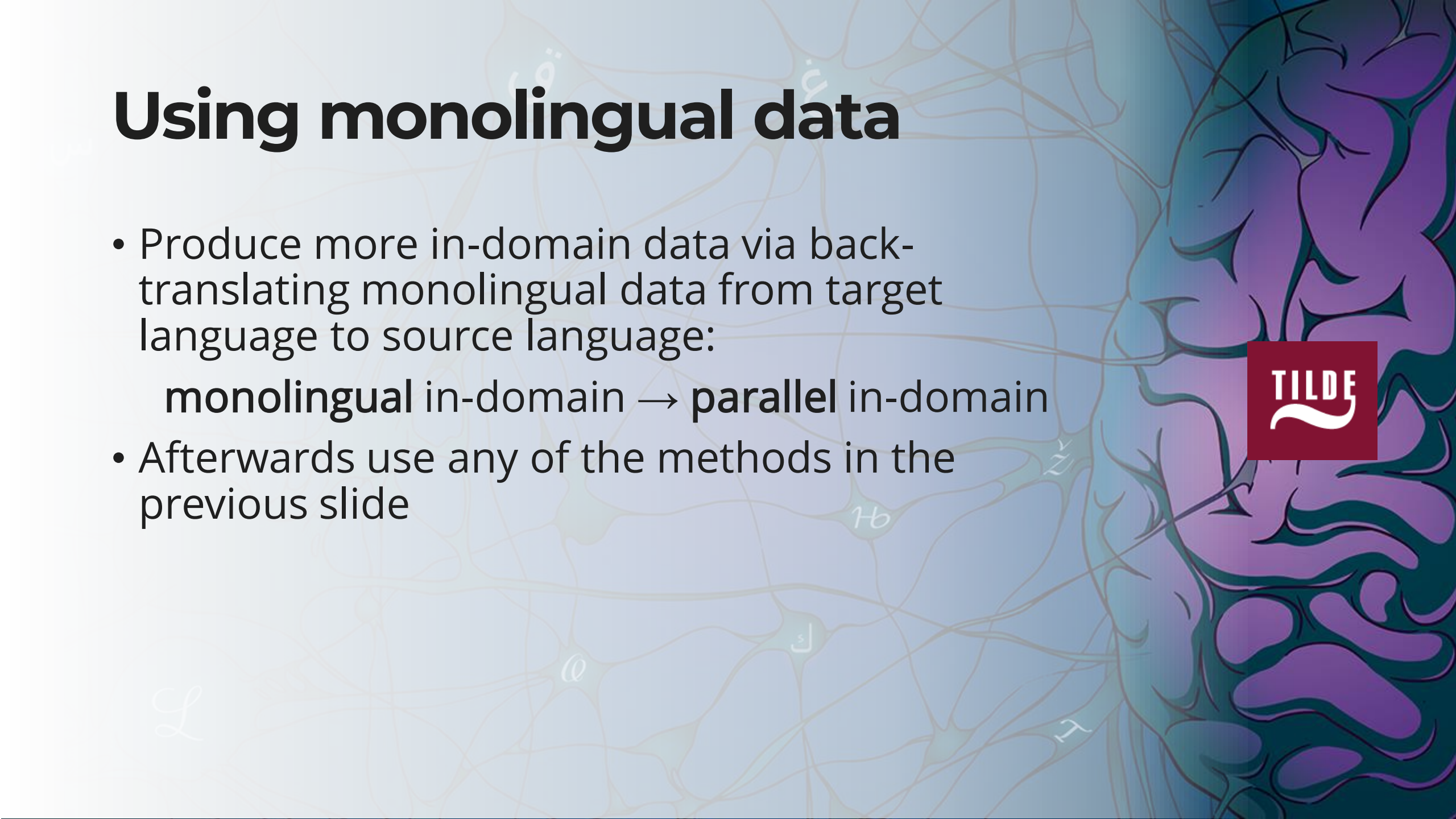


Using monolingual data

- Produce more in-domain data via back-translating monolingual data from target language to source language:

  monolingual in-domain → **parallel** in-domain

- Afterwards use any of the methods in the previous slide



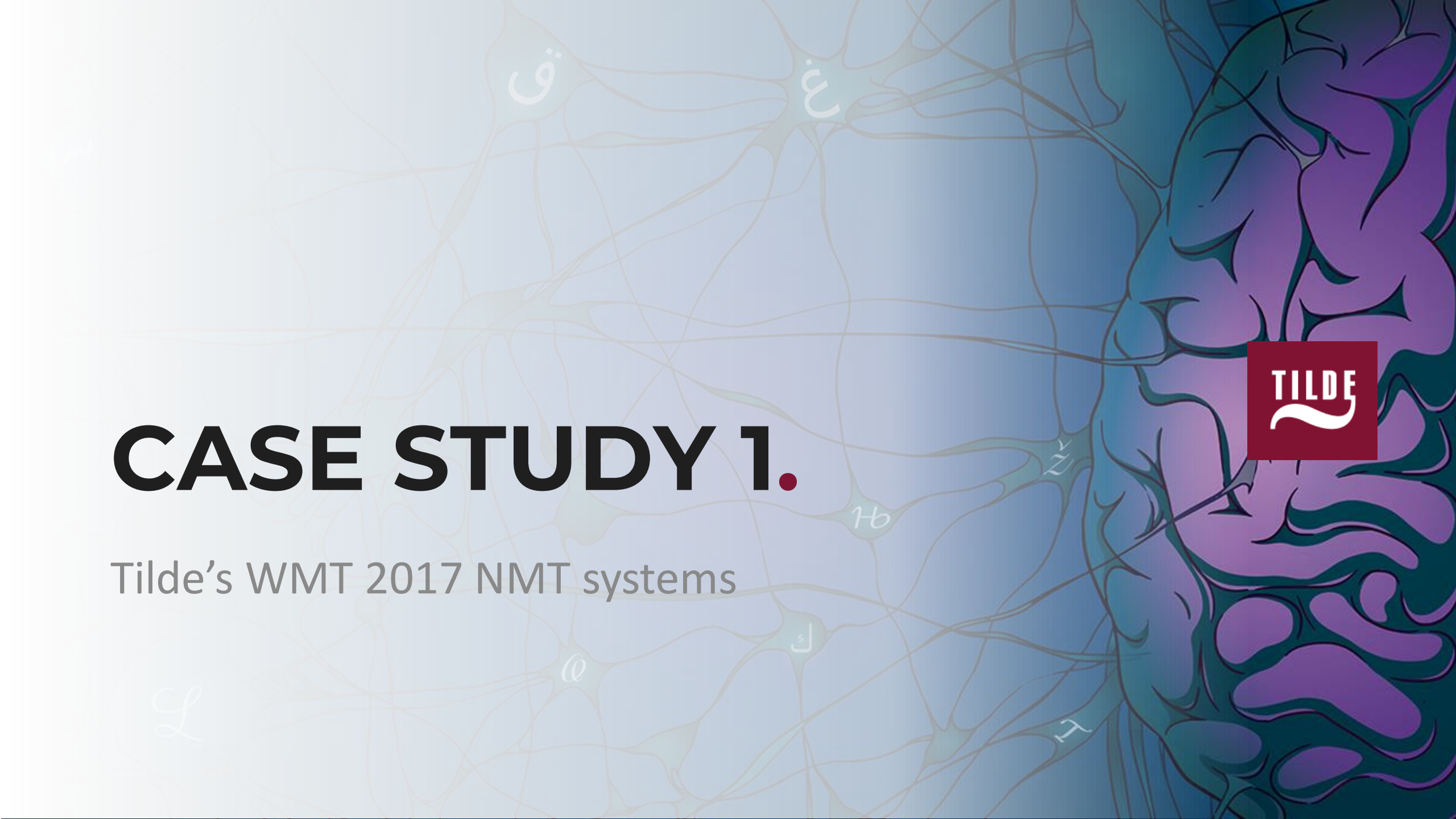
To summarize

- Can an NMT system be trained **on only** a corpora with 5k sentences?
  - No, at least not with current methods.
- What can be done?
  - Identify out-of-domain data. Preferably it is similar.
  - Identify monolingual data in the same domain. It's usually a lot simpler to acquire.
  - Can't just throw it all together – the training process must be structured accordingly. Depends on the size of gathered data.



CASE STUDY 1.

Tilde’s WMT 2017 NMT systems

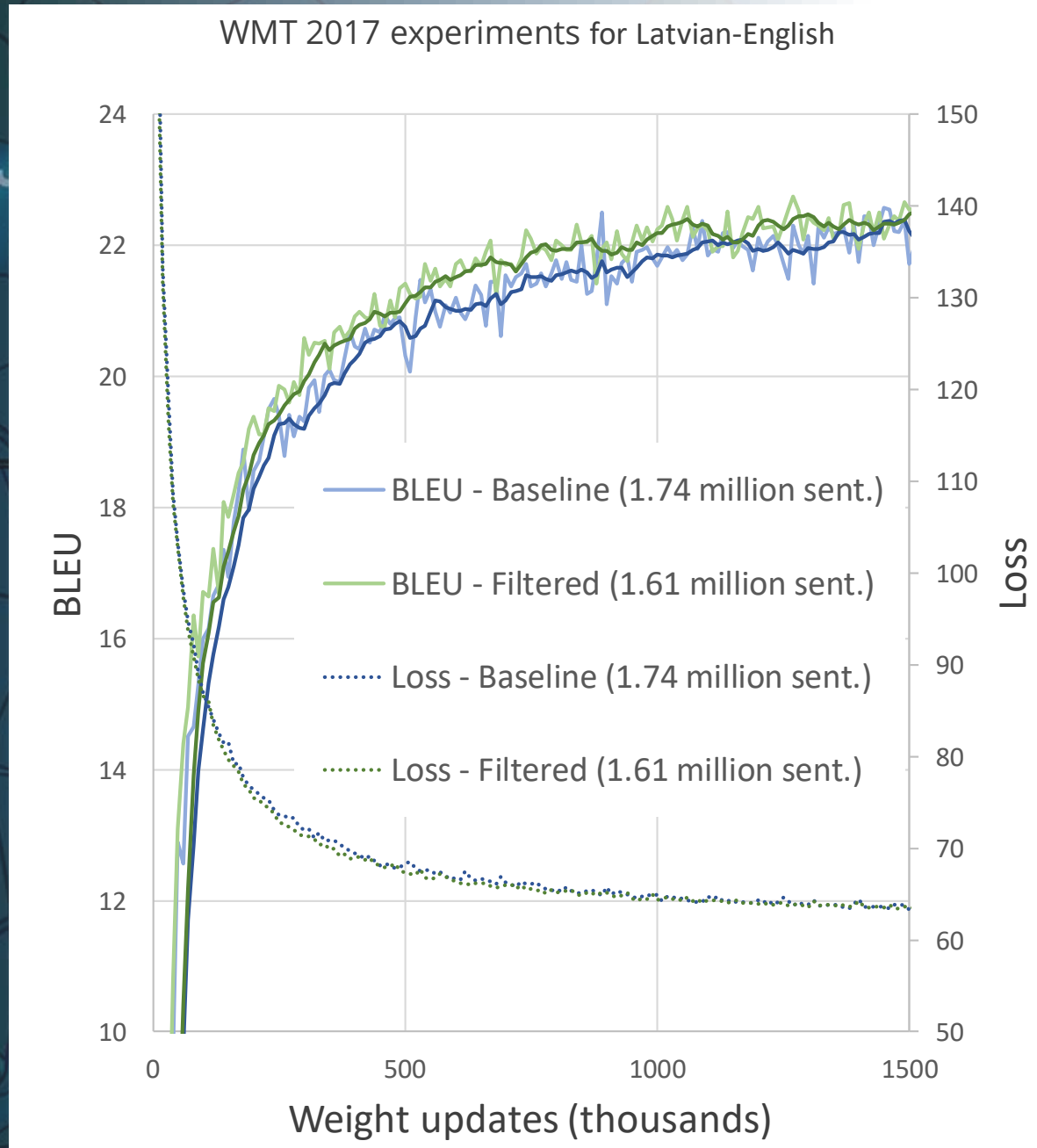


WMT 2017

- WMT is an annual workshop/conference featuring multiple machine translation tasks.

- News translation task – English-Latvian and Latvian-English. What is given?
  - 4.5M parallel out-of-domain sentences (general domain);
  - 38M/370M monolingual in-domain sentences (news);
  - 2k sentence in-domain development set;
  - **No parallel in-domain data.**





DATA CLEANING.

- Data cleaning allows to train the same or better quality NMT systems with less data.
- For SMT systems, data filters tend to lower MT system quality. For NMT, the opposite is true.
- The graph shows scores after the generic filters (from Moses SMT) were applied.

DATA CLEANING.

Corrupt data filters

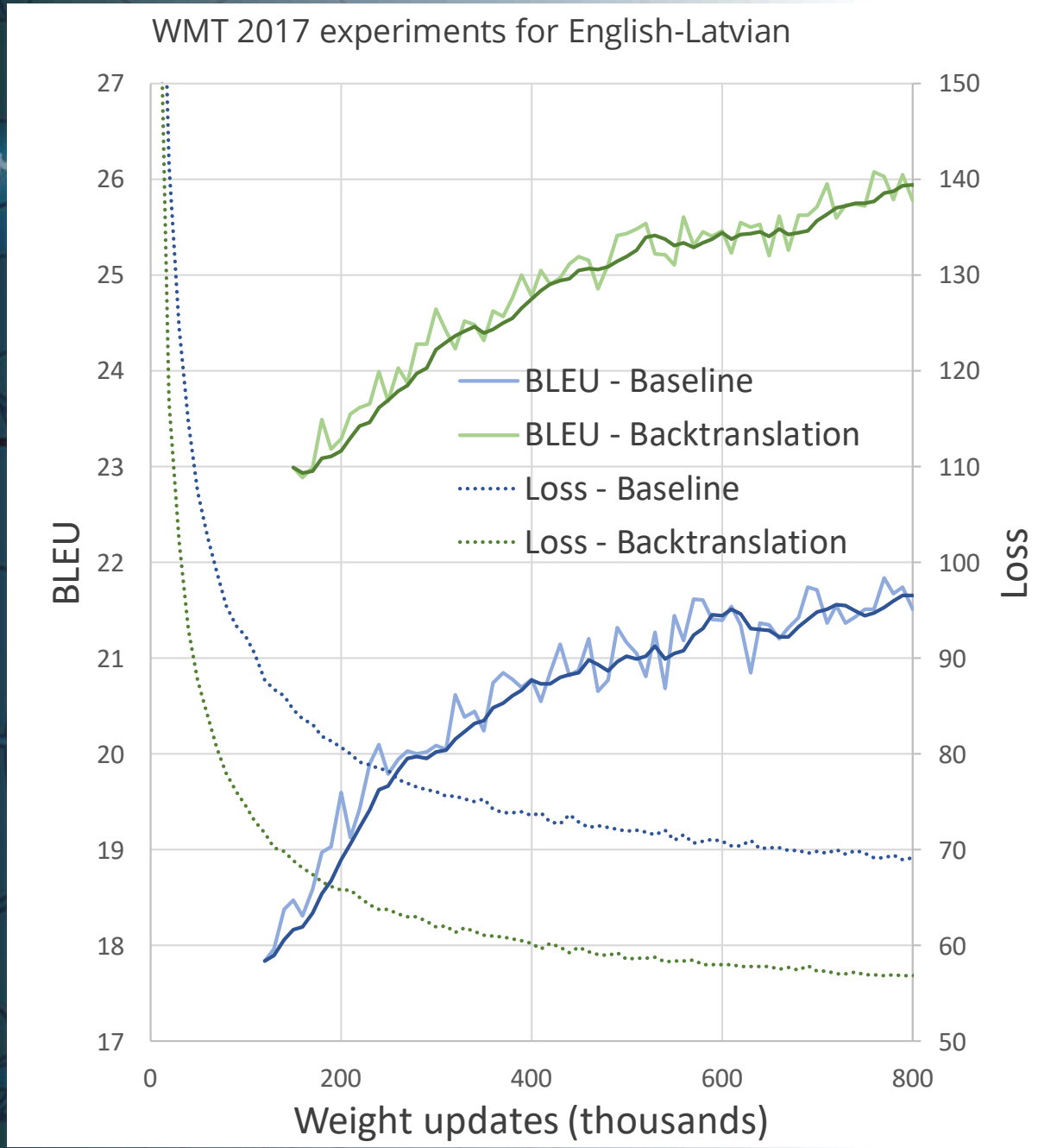
- Corruption through encoding
- Corruption through optical character recognition
- Corruption through (external) data processing workflows
- Non-textual data filtering

Non-parallel data filters

- Incorrect language detection
- Low content overlap filtering
- Length ratio filtering

Data normalisation

- Escaped symbol decoding
- Control symbol removal/replacement
- Tag removal
- Punctuation normalisation



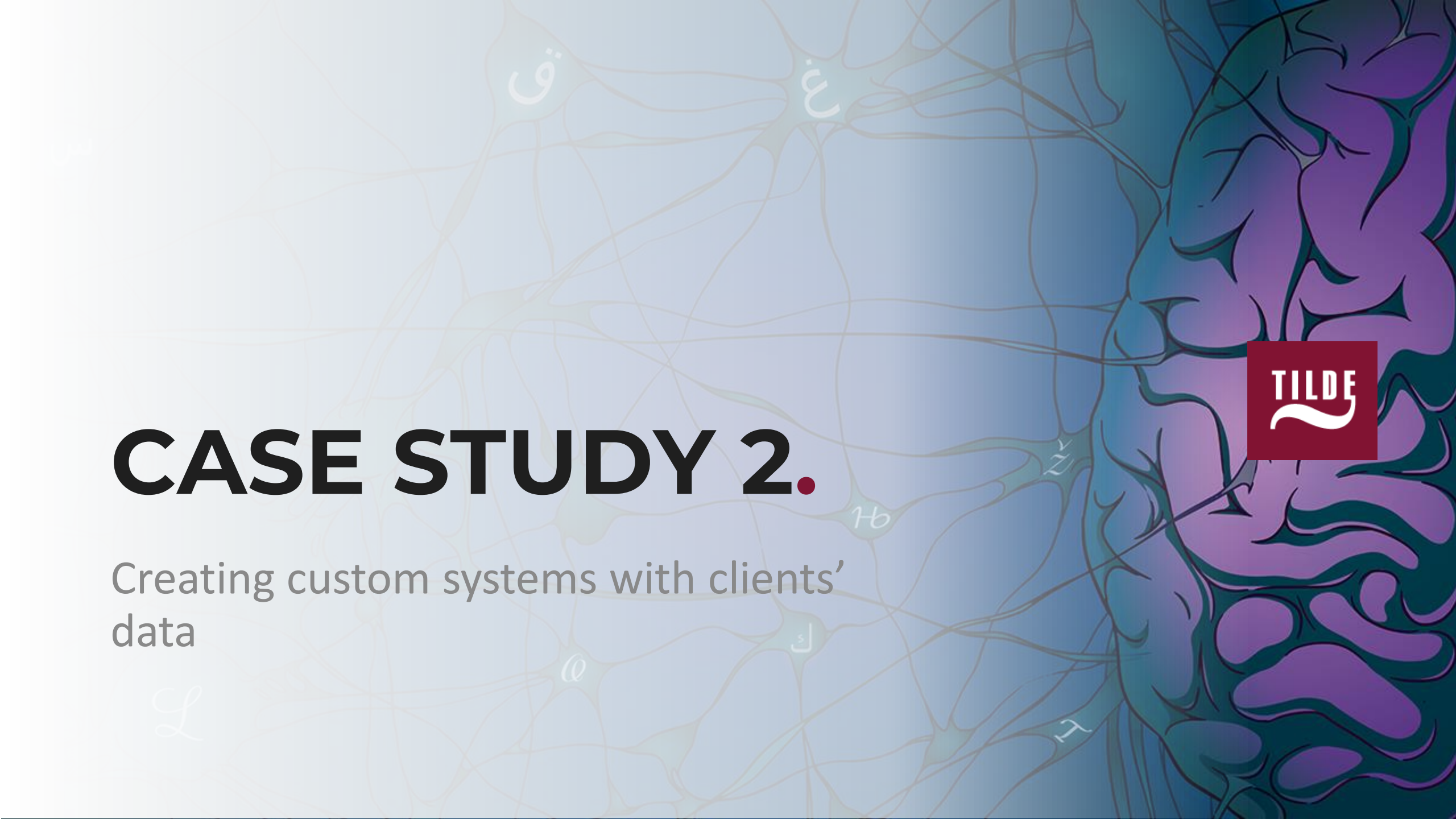
BACK-TRANSLATION

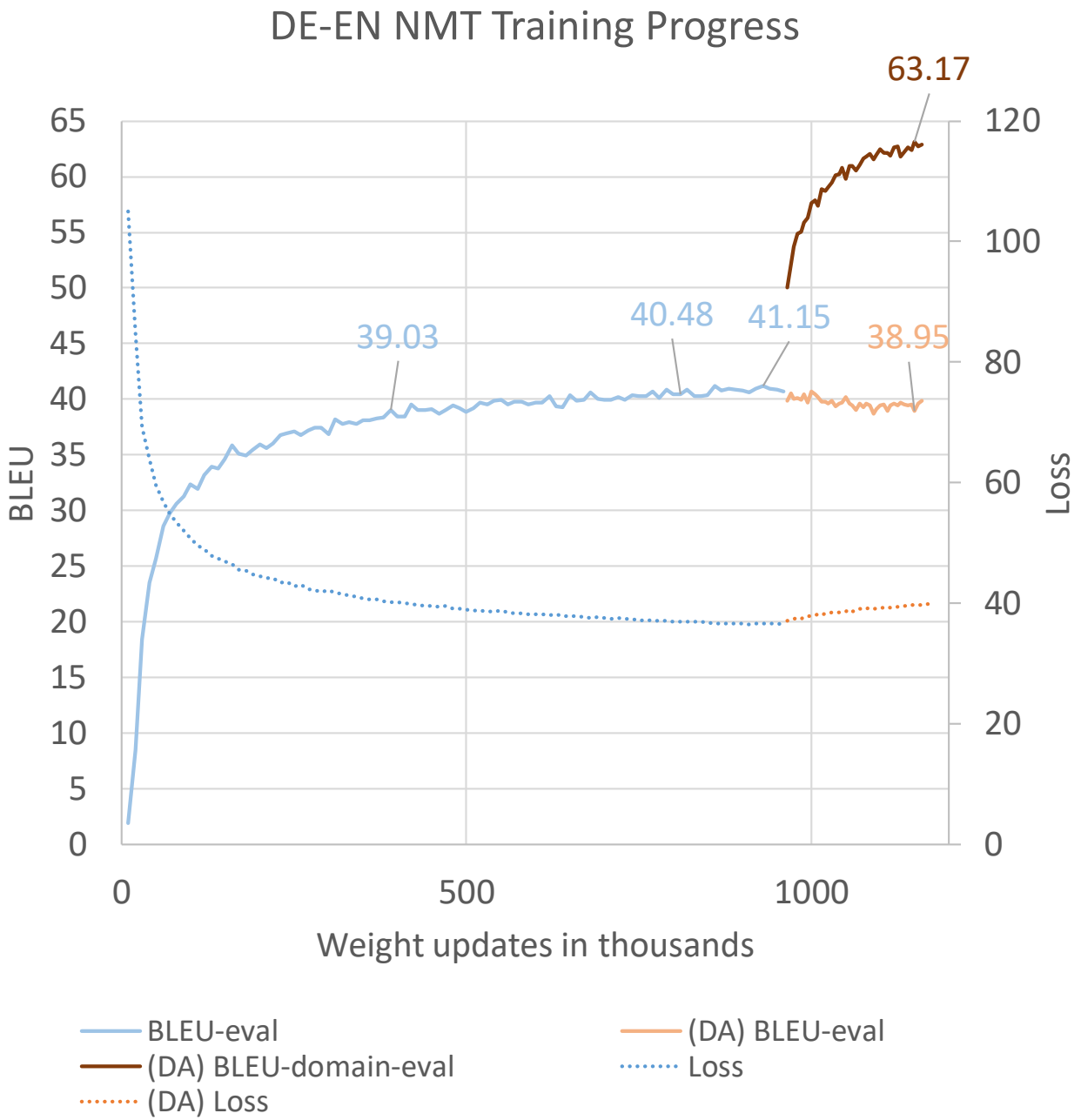
Not all back-translated data are correct, but the models are able to learn from such data because:

1) the target is correct,
2) the data is from our required domain,
3) our original training data is from a broader domain.

CASE STUDY 2.

Creating custom systems with clients' data





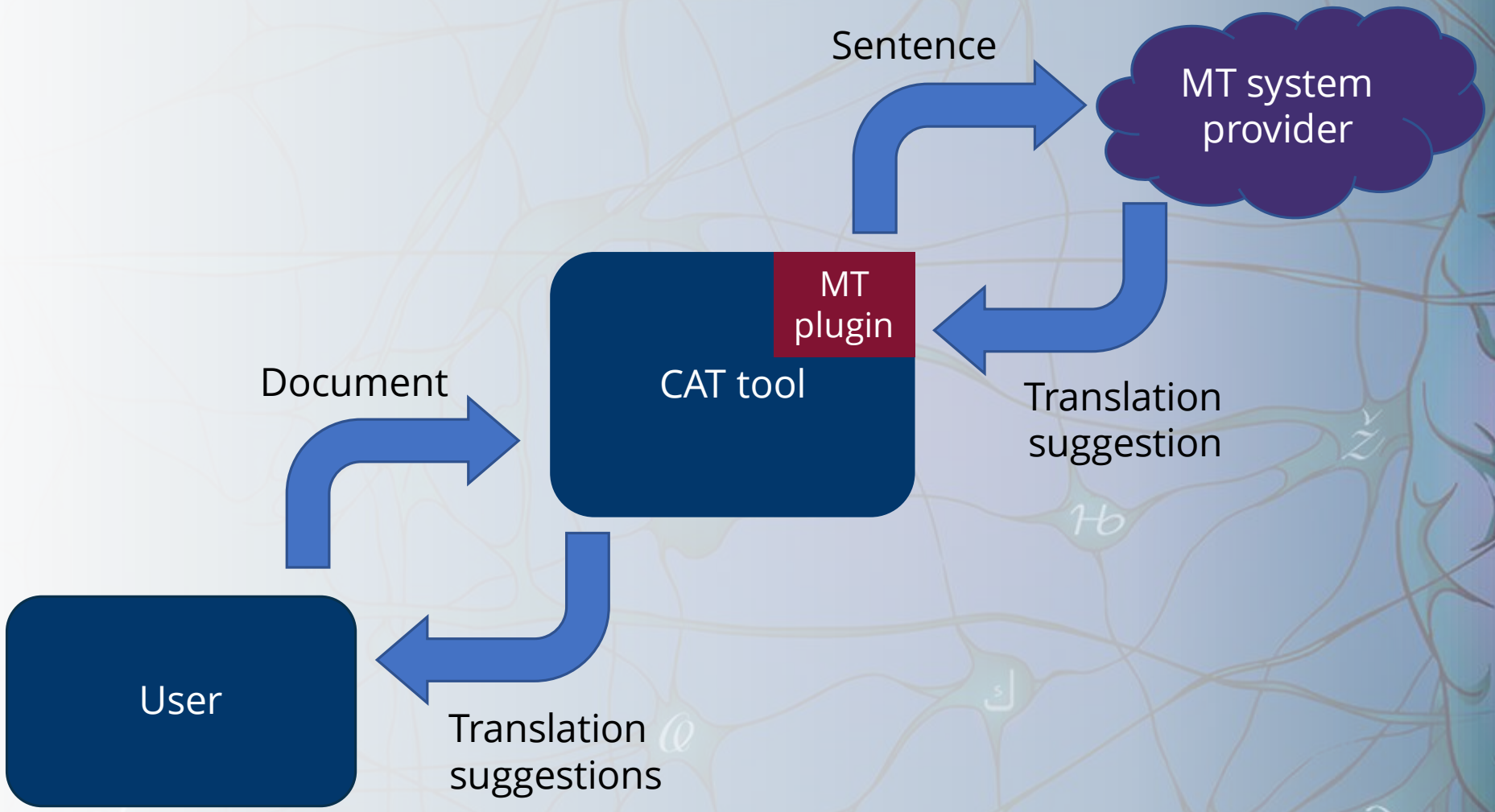
DOMAIN ADAPTATION

- Start out with a large general domain data-set.
- After the model has converged, continue training on a smaller domain-specific data set.

MT IN CAT TOOLS.



MT workflow


Sentence
MT system provider
MT plugin
CAT tool
Document
Translation suggestion
User
Translation suggestions

Plugins are CAT tool specific

- A dedicated plugin must be developed to support a specific CAT tool.
- Different ways to provide the plugin to the end-user:
  - Either already shipped with the CAT tool (e.g. Memsource, MemoQ, MateCat);
  - Or installed on the user's computer (SDL Trados Studio).
- All provide same basic functionality – MT suggestions or pre-translation.
- Advanced functionality might differ – tag support, MT system selection capabilities.

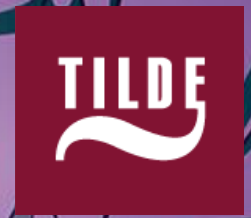


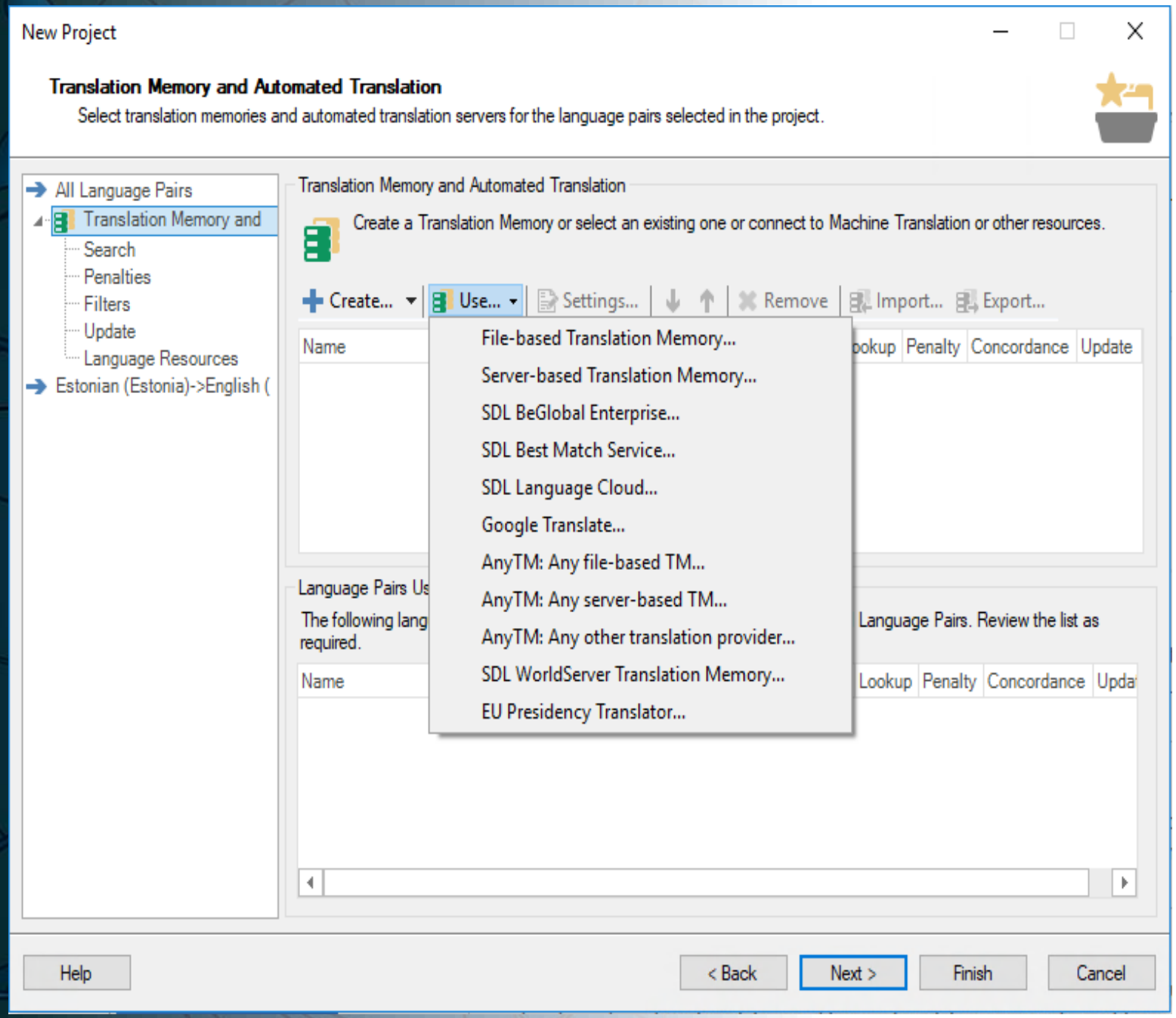
Role of an MT system provider

- A dedicated infrastructure is necessary for MT system hosting:
  - Either in the cloud,
  - Or on premise.
- Provides interface for MT system training and/or trains MT systems and provides them to the user.
- Provides an API endpoint used by the MT plugins in CAT tools.
- Other features – data-set repository, user management, document translation in the web, etc.

NMT-specific limitations

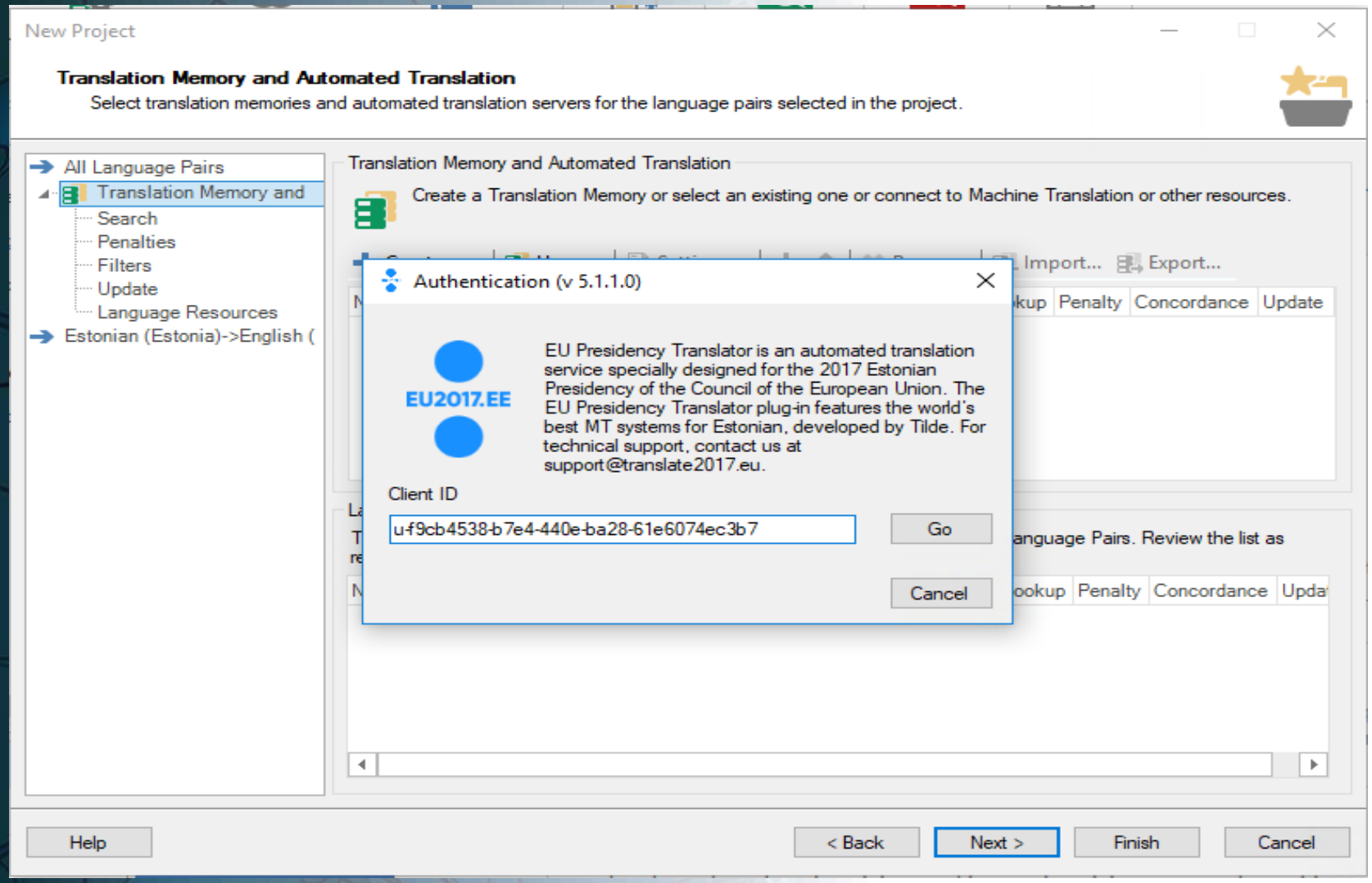
- To enable tag support (essential for formatted documents):
  - NMT model architecture must support alignment extraction,
  - NMT systems must be trained to deal with unknown elements in text.
- Using terminology in NMT is an open research problem:
  - It is best to use a dedicated Terminology plugin to assist the translator.

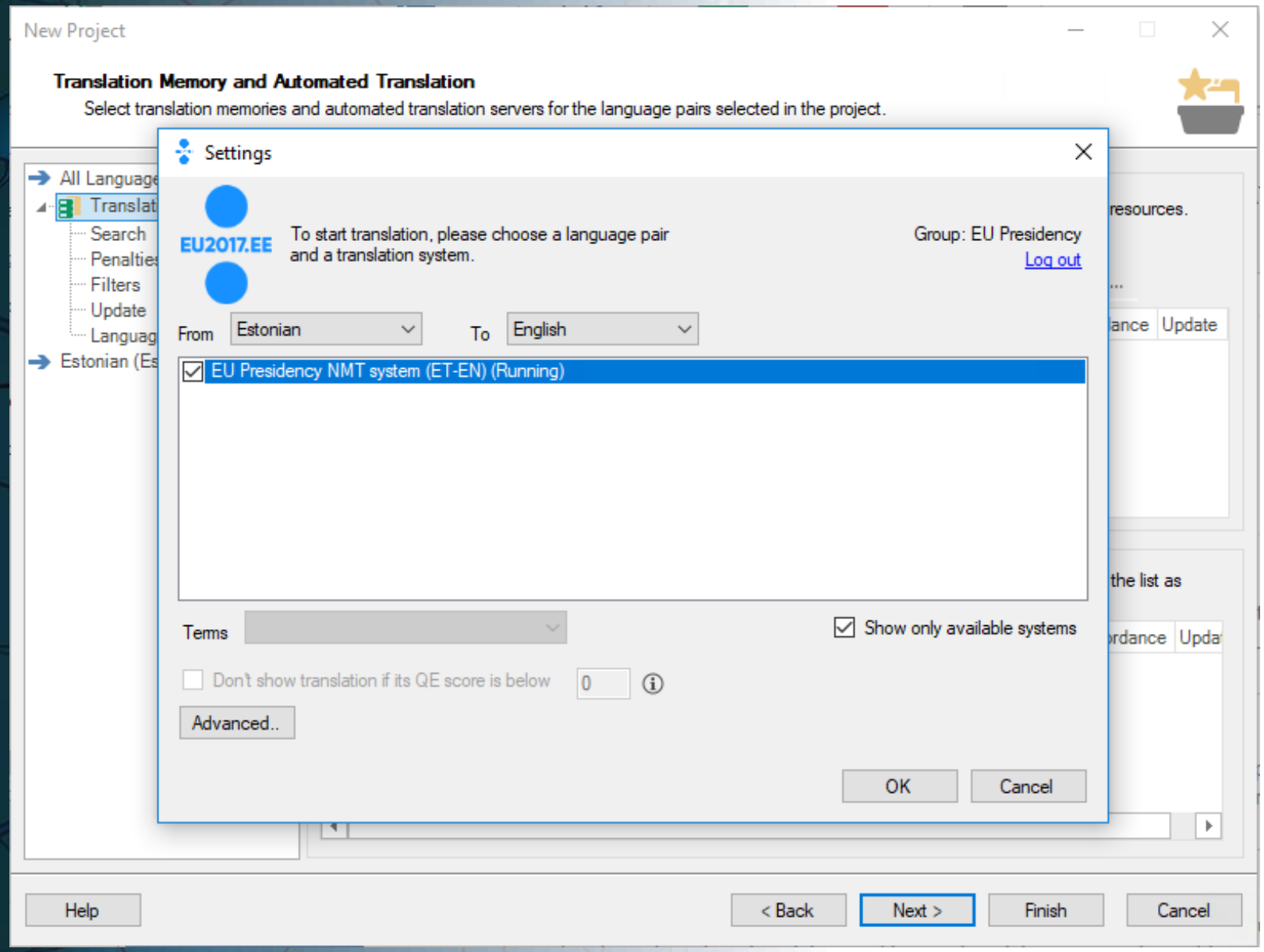




PLUGIN SCREENSHOTS

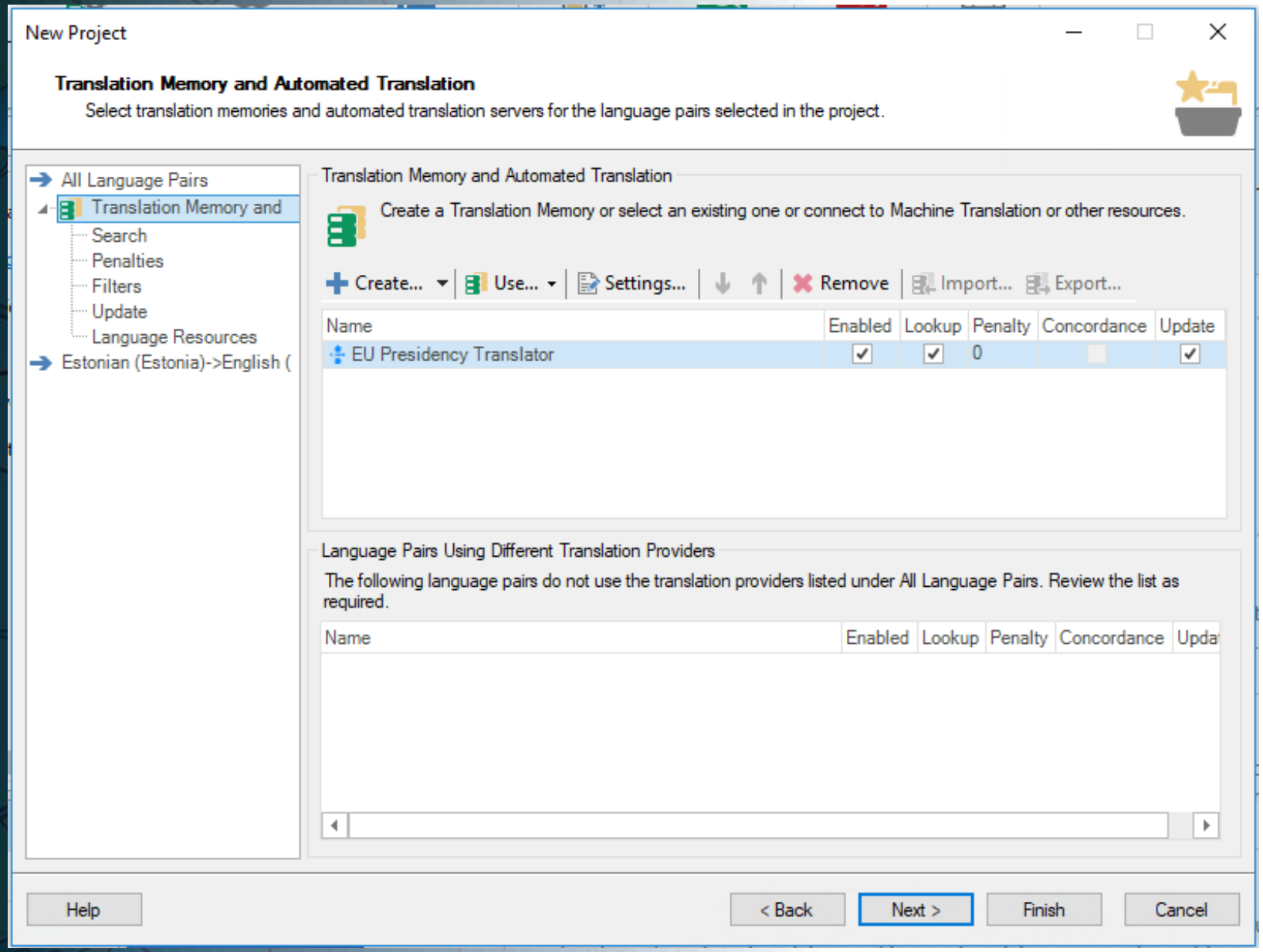
PLUGIN SCREENSHOTS



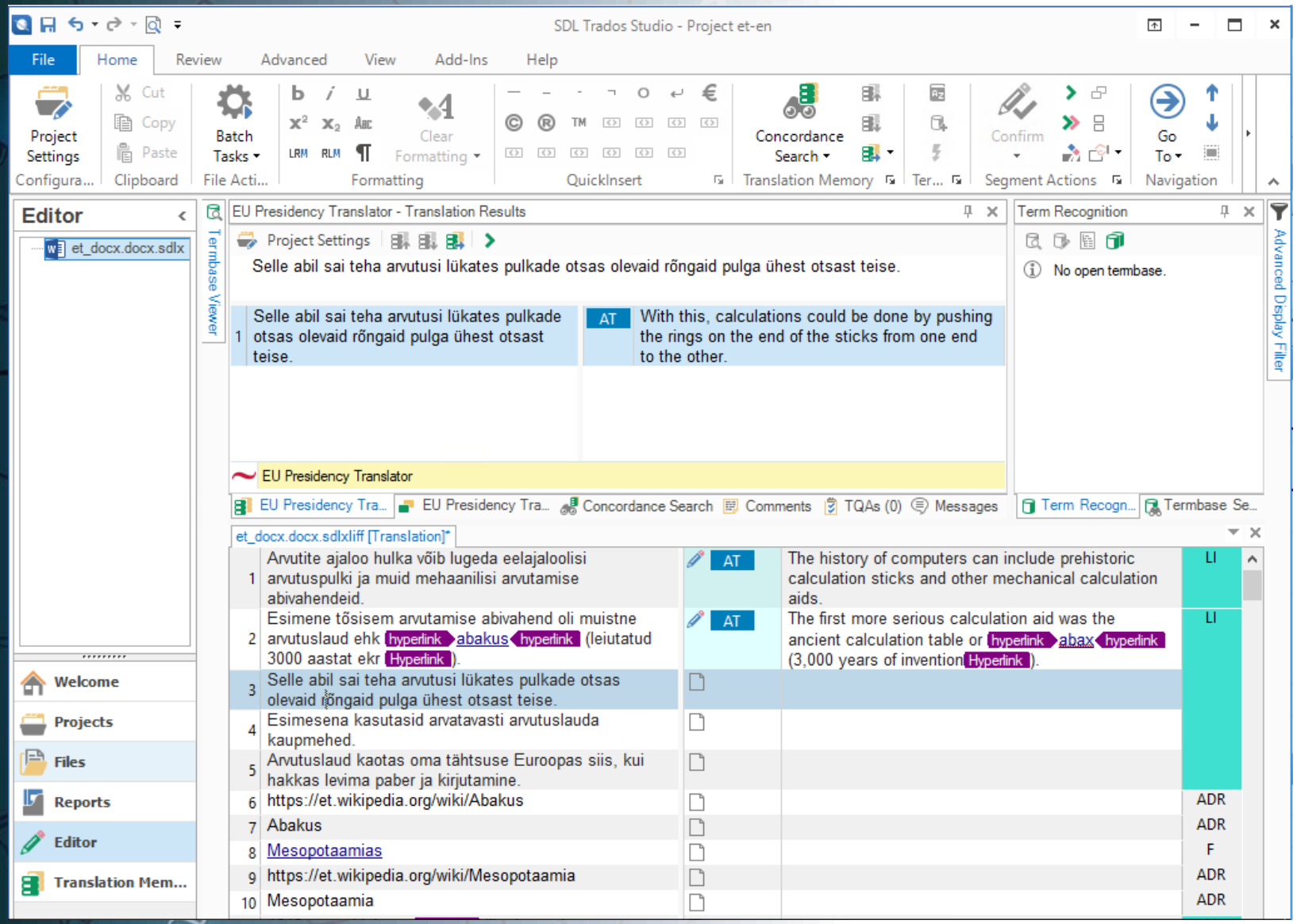


PLUGIN SCREENSHOTS

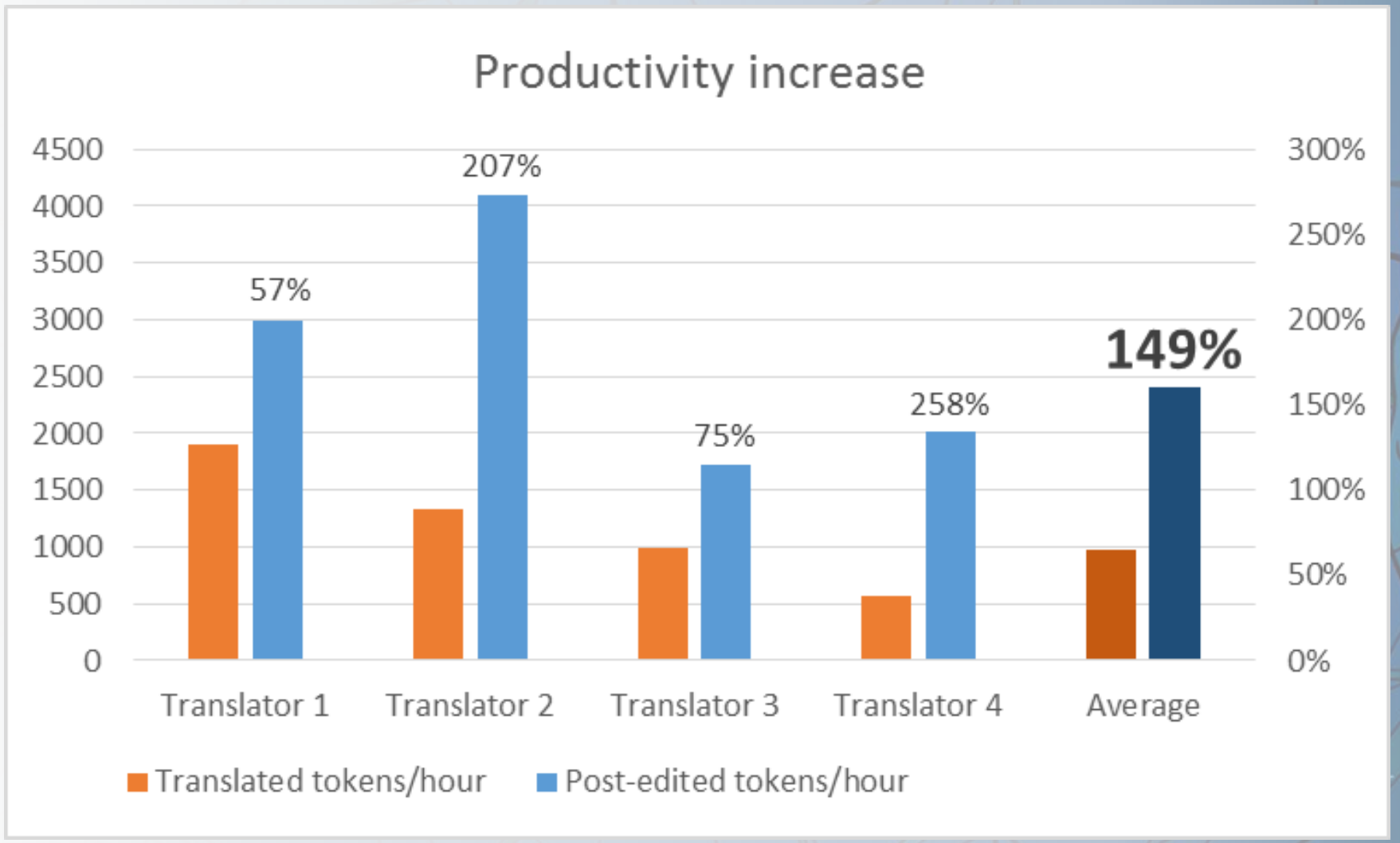
PLUGIN SCREENSHOTS



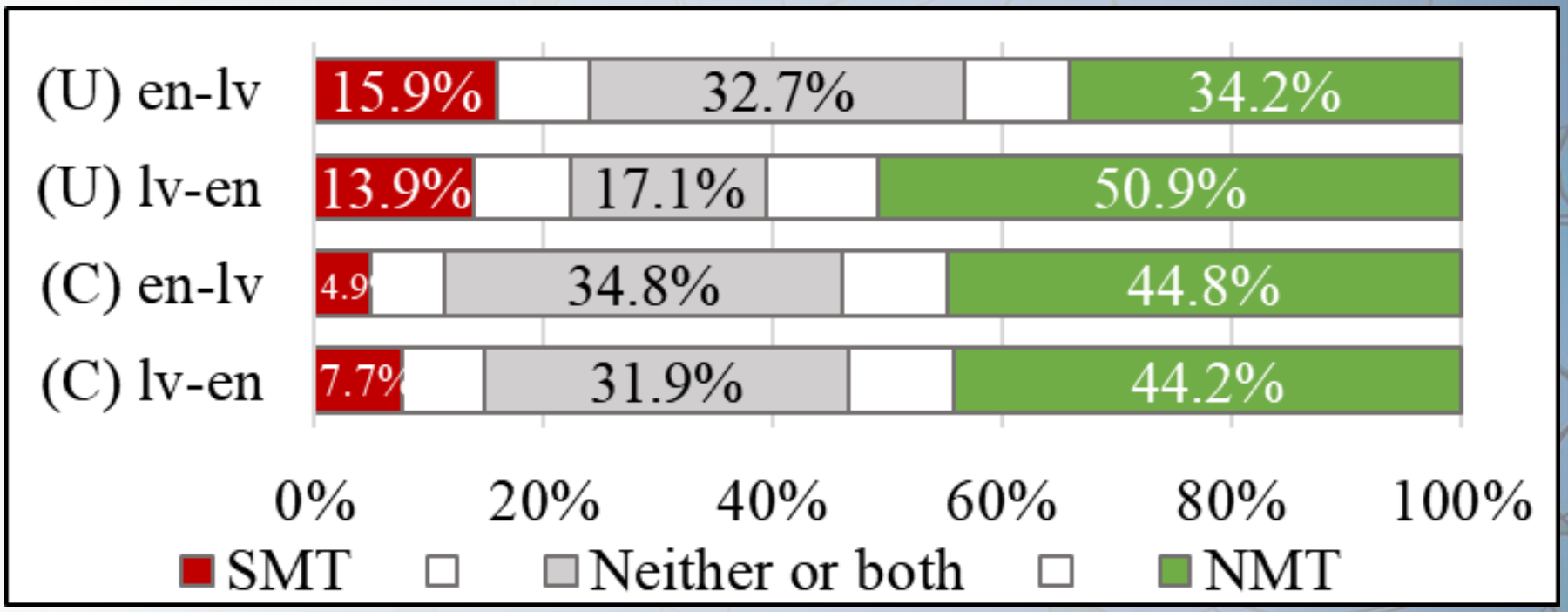
PLUGIN SCREENSHOTS



Productivity gains with MT



SMT vs. Neural MT (human evaluation)



Human evaluation: NMT is better

REFERENCES.

The latest scientific publications on NMT by Tilde are:

Pinnis, M., Krišlauks, R., Miks, T., Deksne, D., & Šics, V. (2017). **Tilde's Machine Translation Systems for WMT 2017**. In *Proceedings of the Second Conference on Machine Translation (WMT 2017), Volume 2: Shared Task Papers* (pp. 374–381). Copenhagen, Denmark: Association for Computational Linguistics.

Pinnis, M., Krišlauks, R., Deksne, D., & Miks, T. (2017). **Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data**. In *Proceedings of the 20th International Conference of Text, Speech and Dialogue (TSD2017)*. Prague, Czechia.

Pinnis, M., Krišlauks, R., Deksne, D., & Miks, T. (2017). **Evaluation of Neural Machine Translation for Highly Inflected and Small Languages**. In *Proceedings of the 18th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2017)*. Budapest, Hungary. *In press.*


Thank you!
TILDE