



LOCALIZATION WORLD
CONFERENCE & EXHIBITS

29 Oct 2014

Up-to-date MT with Product-Specific Terminology

Ventsislav Zhechev

ventsislav.zhechev@autodesk.com

Autodesk Development Sàrl, Neuchâtel, Switzerland



Rule-based vs. Statistical

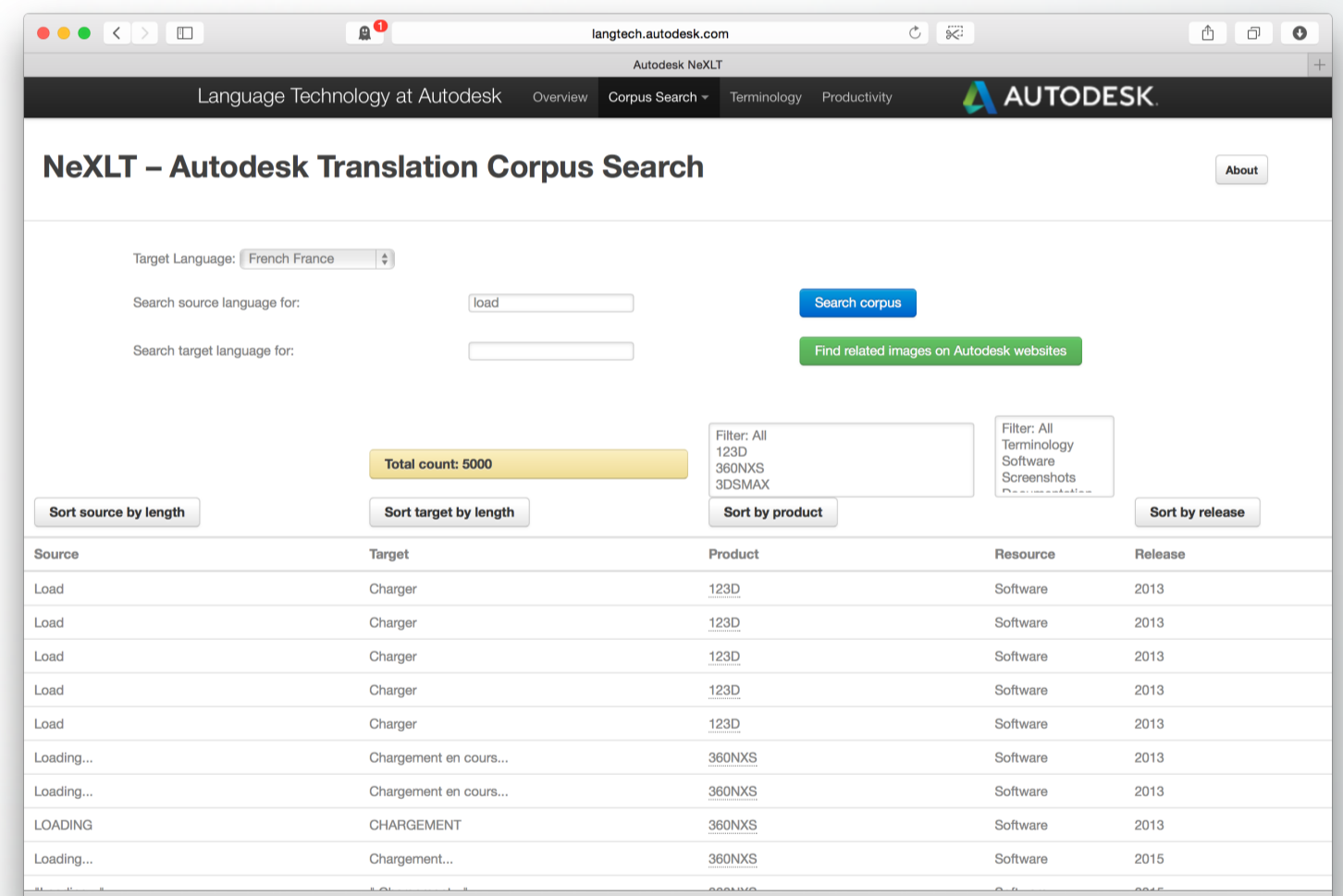
- Rule-based MT is significantly impaired when processing new content
 - pre-processing new terminology is a must
 - caveat:
 - hybrid systems with statistical post-processing
- Statistical MT is expected to robustly handle new content
 - still, it can only handle with confidence text similar to what it was trained on
 - pre-processing new terminology is beneficial

Legacy Glossaries

- Manually created and maintained glossaries
- Individual glossaries for different geographical language groups
 - FIGS, EEUR, APAC
 - different levels of coverage
- Conflicting translations
- Built by humans for humans
 - comments inline with terms
- Untranslated terms

NeXLT

- Enables translators to query all segments that have been translated/post-edited at Autodesk to date
- Main use: provide guidance when translators are unsure about how to translate a particular term or phrase
- Search results are sortable and filterable by product (ACD, REVIT, ...) and origin (software, documentation)

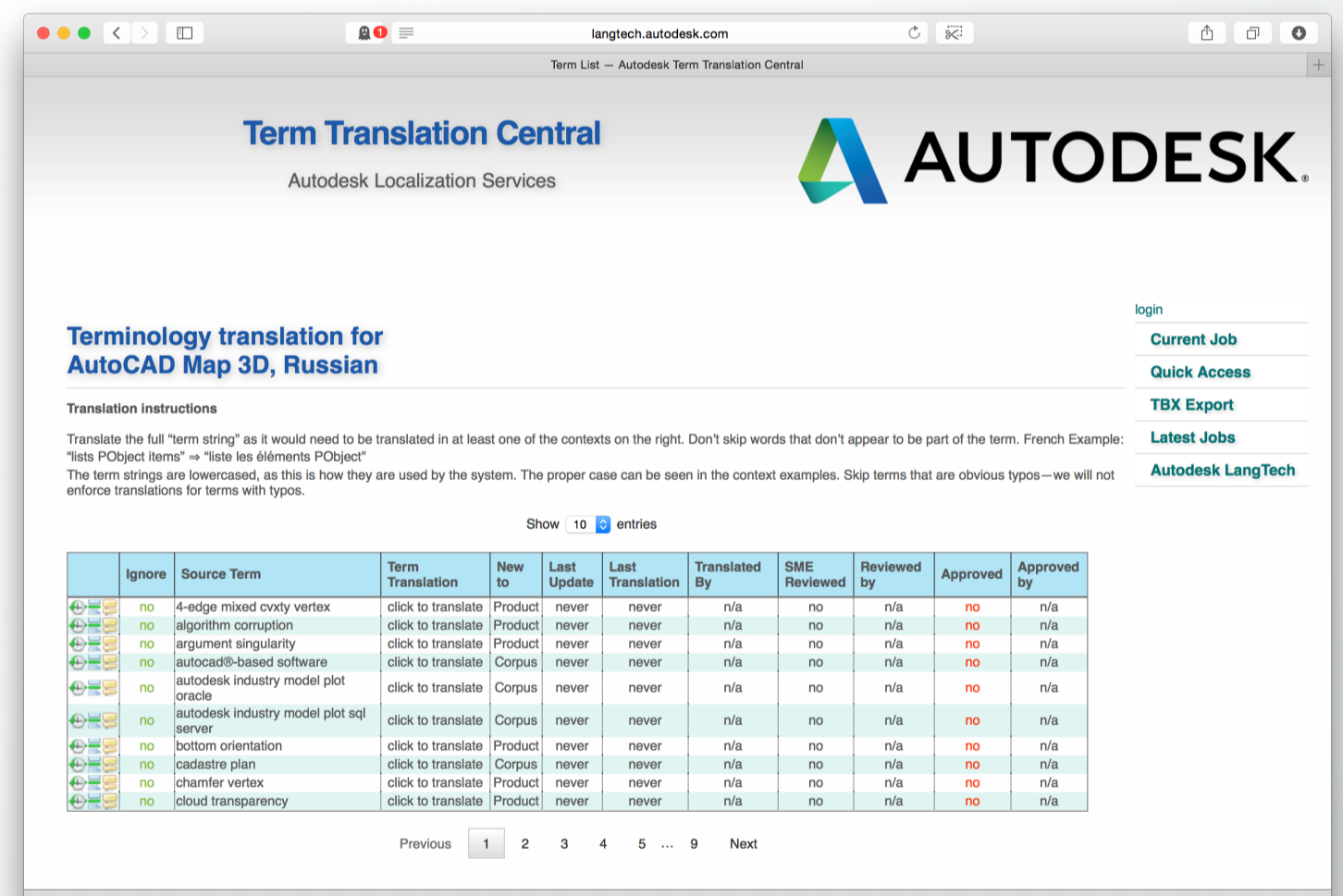


Handling New Content

- When a new feature is developed for one of our products
 - new UI strings are generated
 - technical writers document the feature producing new documentation content
 - the new content (UI and documentation) is translated
- This new content may contain terminology related to the new product feature that is yet unknown
 - to our MT engines
 - to our translators

Term Translation Central

- Identify terms that haven't been translated yet in new content
- Simplifying assumption: Every noun phrase is a term
- New content is submitted to the Term Translation Central as soon as it is available
- Previously unknown NPs are extracted: they can be
 - new to all Autodesk content; or
 - new to this product's content



The screenshot shows the Autodesk Term Translation Central interface. The page title is "Term Translation Central" with the Autodesk logo. The current job is "Terminology translation for AutoCAD Map 3D, Russian". The interface includes a "login" link, "Current Job", "Quick Access", "TBX Export", "Latest Jobs", and "Autodesk LangTech" links. Below the navigation is a "Translation instructions" section with a "Show 10 entries" dropdown. The main content is a table with the following columns: Ignore, Source Term, Term Translation, New to, Last Update, Last Translation, Translated By, SME Reviewed, Reviewed by, Approved, and Approved by. The table contains 10 rows of data, each with a "click to translate" link in the "Term Translation" column.

Ignore	Source Term	Term Translation	New to	Last Update	Last Translation	Translated By	SME Reviewed	Reviewed by	Approved	Approved by
<input type="checkbox"/>	4-edge mixed cvxty vertex	click to translate	Product	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	algorithm corruption	click to translate	Product	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	argument singularity	click to translate	Product	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	autocad®-based software	click to translate	Corpus	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	autodesk industry model plot oracle	click to translate	Corpus	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	autodesk industry model plot sql server	click to translate	Corpus	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	bottom orientation	click to translate	Product	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	cadastre plan	click to translate	Corpus	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	chamfer vertex	click to translate	Product	never	never	n/a	no	n/a	no	n/a
<input type="checkbox"/>	cloud transparency	click to translate	Product	never	never	n/a	no	n/a	no	n/a

Term Identification

1. New translation job comes in

start automatic loads generation .

2. Part-of-speech tagging

VB JJ NNS NN .

3. Chunking

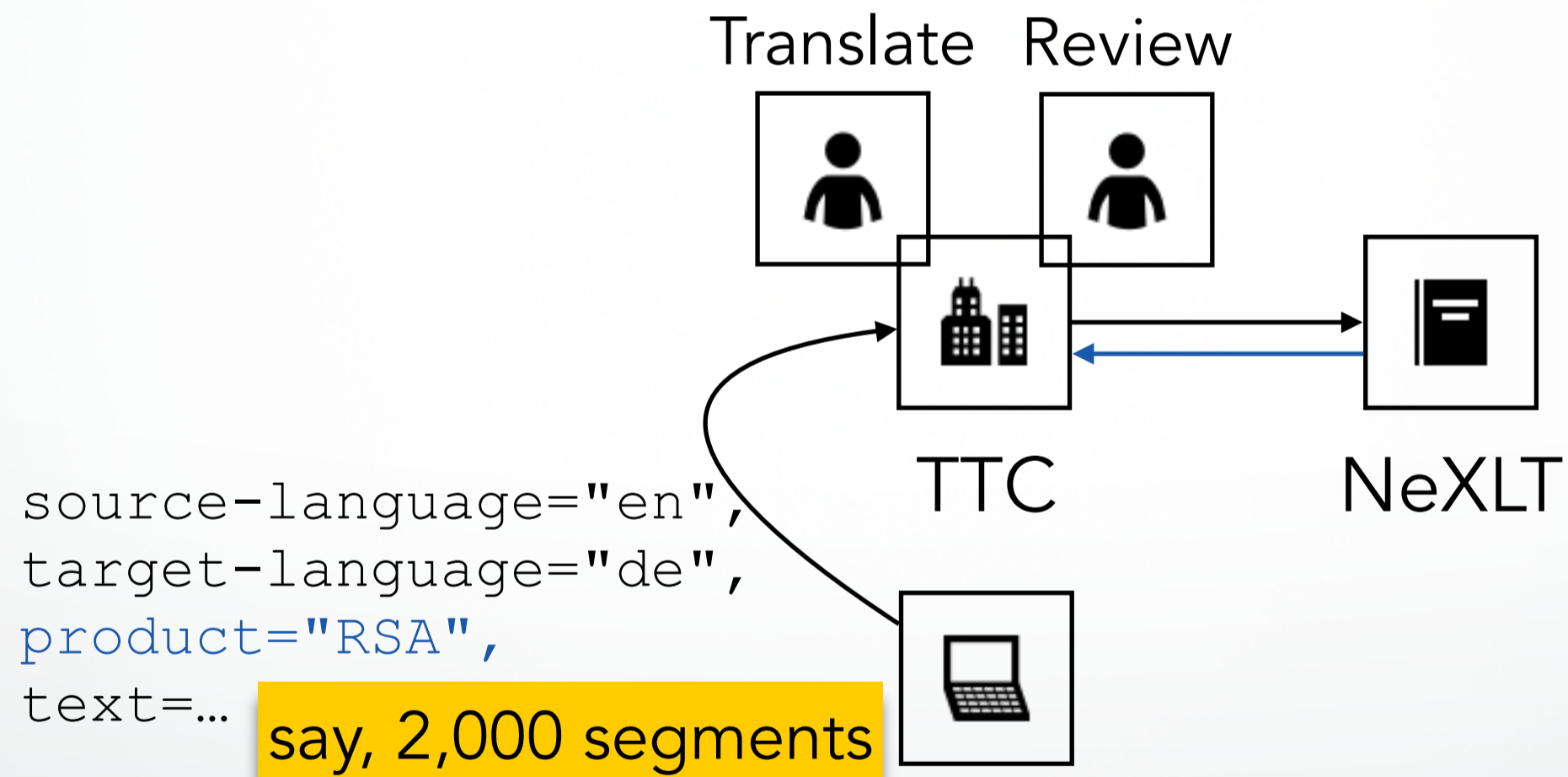
CHUNK: { (<NN* | VBN>*) (<JJ* | VBN>*) (<NN* | VBN>) (<NN*>+) }

4. Manual Translation

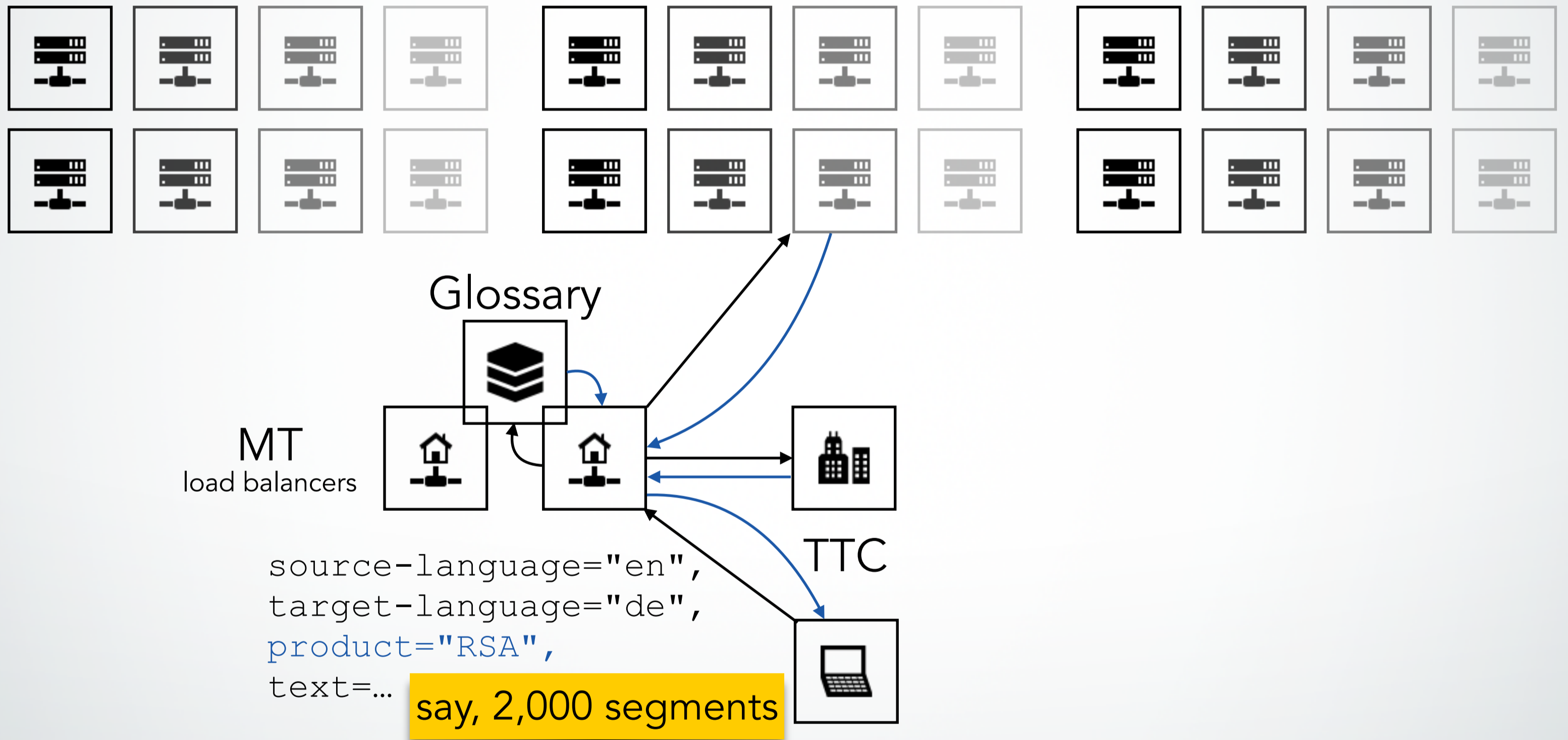
automatische Lastengenerierung

5. Store translation and make it available to MT engines

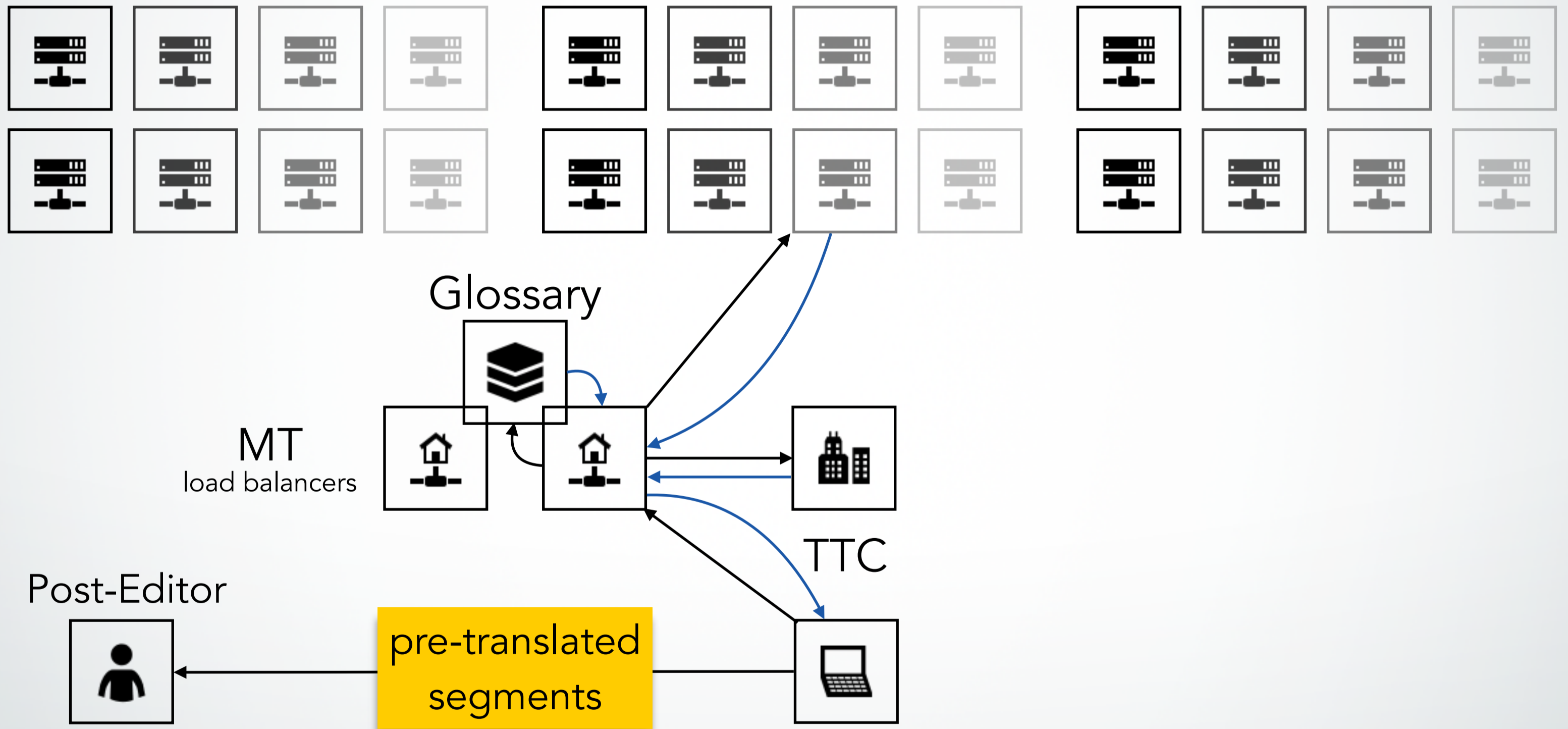
Term Translation (Step 1)



Machine Translation (Step 2)



Post-Editing (Step 3)



Process Benefits

- MT output contains pre-audited translations for new content, per product/language combination
 - which helps improve consistency
 - but not necessarily reduce post-editing
- Translator pain is reduced
 - need to consult reference materials less frequently
 - but still need to correct morphological agreement

To Conclude

- Your MT is always up-to-date and ready for processing new content
- You have the added benefit of your MT producing the correct terminology for each of your products, without involved domain-adaptation work

