Efficient Translation Production for the Multilingual Web

Matthias Heyn, SDL

VP Global Solutions

Agenda

“Understand current state of art to increase translator productivity”

- **Translation Production**
- **Productivity Accelerators**
- **Standards**
- **Current Trends**

Translation Production

Content is either …

- **Translated by professional translator**
- **Or, the "occasional" translator**
  - Non-linguist, Subject matter specialist (reviewer), Crowd sourced, …
- **Or, left un-translated**
  - Not relevant, too costly, too much overhead involved, …

This presentation focuses on content produced by professional translators

Which in turn are handling increasing volumes of Web Content

Productivity Environments

- **Today, content workers utilize specialized productivity environment(s)**

Content Worker	Application Class	Prominent Example
Graphic Designers	Graphic tools	Adobe Photoshop
Audio Producers Musicians	DAW (Digital Audio Workstation)	Steinberg Cubase
Architects	3D modeling program	Google Sketch up
Engineers	CAD (Computer Aided Design)	Autodesk AutoCAD
Game Developer	Game Engine	Epic Games Unreal Engine
Translators	CAT (Computer Aided Translation)	SDL TRADOS TWB / SDL Studio

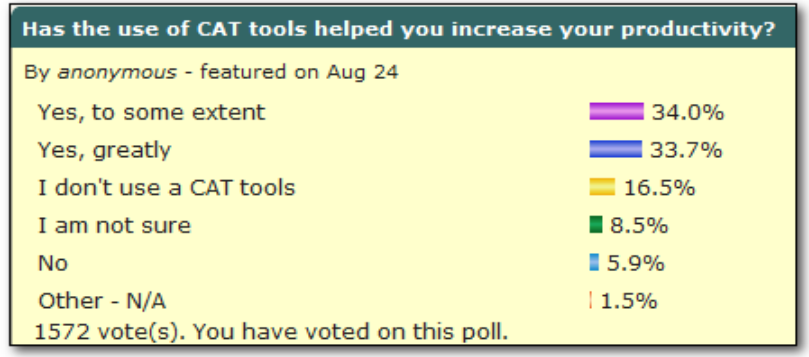
Translation Editor at the core of CAT

Professional Translation can be done …

- **In principle, in any authoring editor (desktop/browser)**
  - However, with limited productivity (in the range 800-1500 words per day) and high efforts maintaining consistency and accuracy.
- **Using Microsoft Word + Plug-ins**
  - Plug-in to translation productivity tool
  - Hard dealing with structured content
- **Using a Dedicated Translation Editor**
  - Depending on various factors: productivity boost in the range 2000 to 5000 words per day
  - Well established market for professionals

Translation Editor Specifics

- **Explicit representation of source and target language**
  - Side-by side or top down
- **Segmentation**
- **Abstraction of formatting information**
  - Allows for same editing environment for any input format (“learn one tool, translate anything”)
- **(Dynamic) preview of formatted text**
- **Selection of “Translation providers” to boost productivity**



Public ProZ Poll August 24 reply from 1572 translators
http://www.proz.com/polls/5474

Key Productivity Accelerators

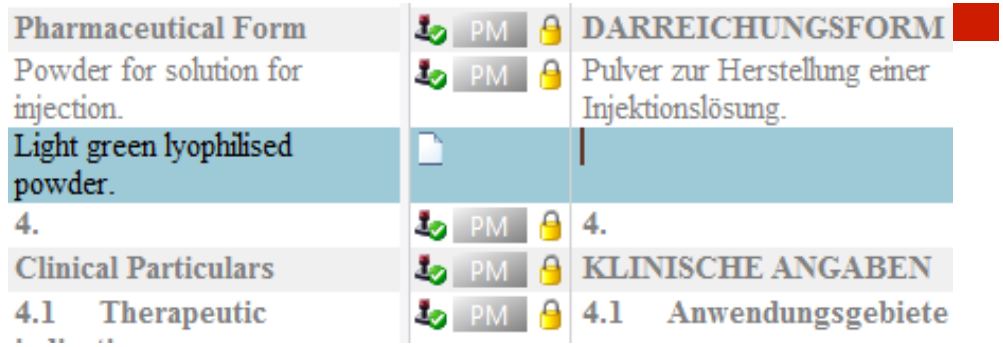
Topic Level document, page, fragment, chunk, …	**Segment Level** sentence, header, footnote, table cell, …	**Subsegment Level** phrase, word, …
Exclusion from translation through markup	Translation Memory	Auto-suggest (dictionary based auto-completions)
“Perfect Matching” utilizing bi-lingual representations	Automated Translation	Placeables, Terms
	Auto-propagation	Concordance

- Impact on effective handling of update translations
- Impact on effective handling of new translations
- Impact on effective handling of document internal redundancies
- Impact on consistency & quality

Topic (Document, …) Level

"Don't translate if it hasn't changed"

(but show it to provide context for the text that has actually changed/ added)

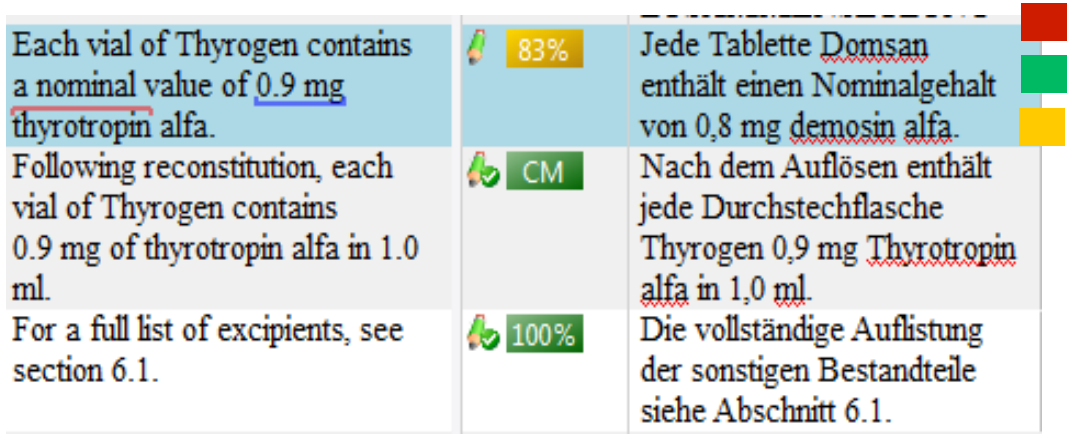


- **Markup exclusions**
  - Use ITS / other convention to lock text
  - Custom arrangements between CMS + Translation System
- **Perfect Matching**
  - Compare text with predecessor translation project and lock what hasn't changed
  - But, high overhead in managing corresponding projects

Significant productivity gains dependent on update frequency

Segment Level : TM

"Don't re-translate if you can reuse an (approved) existing translation"
(but adapt as you need)

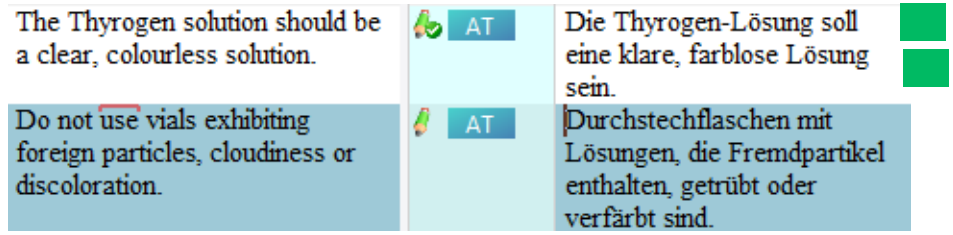


Each vial of Thyrogen contains a nominal value of 0.9 mg thyrotropin alfa.	83%	Jede Tablette Domsan enthält einen Nominalgehalt von 0,8 mg demosin alfa.
Following reconstitution, each vial of Thyrogen contains 0.9 mg of thyrotropin alfa in 1.0 ml.	CM	Nach dem Auflösen enthält jede Durchstechflasche Thyrogen 0,9 mg Thyrotropin alfa in 1,0 ml.
For a full list of excipients, see section 6.1.	100%	Die vollständige Auflistung der sonstigen Bestandteile siehe Abschnitt 6.1.

- **Increasingly sophisticated match type differentiation**
  - 100%, Fuzzies, Context Matches (CM), (ICE)
- **Cascaded TMs, Ranking of TMs**
- **Significant productivity gains dependent on**
  - Availability of relevant TMs
  - Similar content produced again and again

Segment Level : Automated Translations

"Adapt an automated translation proposal"
(instead of translating from scratch)



- **Increasingly accepted by professional translators**
  - Especially using Statistical Machine Translation (SMT)
- **Significant Productivity gains depending on**
  - SMT engine trained with sufficient, relevant (in-domain), high quality (professional translator output) data
  - Translators able to dynamically select "in-domain" trained engine [e.g. "Touchpoints"]
  - Trust scores

Segment Level : Automated Translation

"Trust score" to determine when a proposal is most likely useful and when not

- **Document level**
  - Route documents into alternative production chains (Dynamic Routing)
- **Segment level**
  - Dynamically provide AT proposals

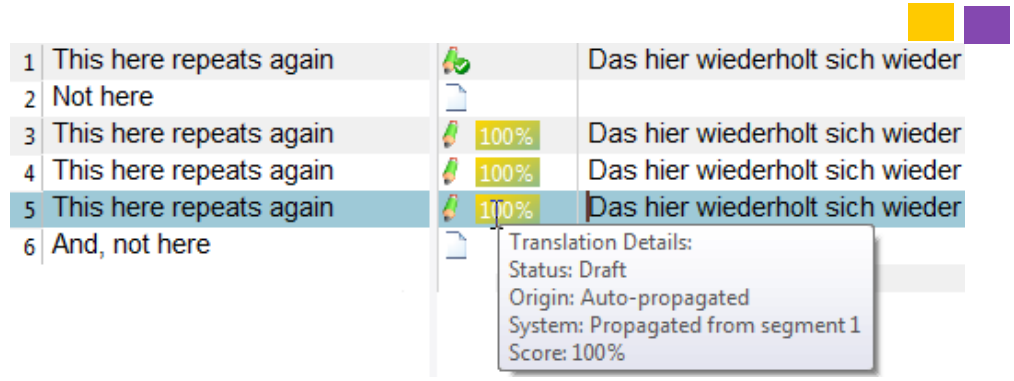
- **TM important to train domain specific SMT engines**
- **SMT important to speed up TM growth**

- **Automate "retraining" of SMT engine / phrase dictionaries in feedback cycle(s)**

Segment Level : Auto-propagation

"Auto-propagate translations for identical source segments"

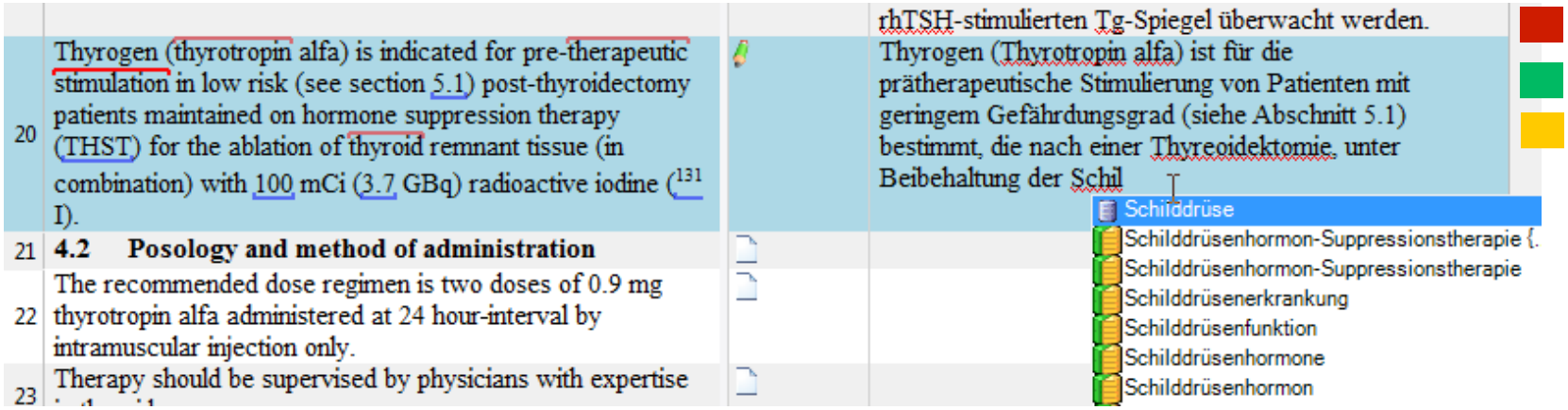
(and ripple through any changes when you change your translation)



- **Productivity gain if text has internal repetitions**
  - Simplifies updating identical segments throughout the content
- **Requires parameters to control behavior**

Subsegment Level : Auto-suggest

"While I type, provide a list of relevant candidates so that I can quickly auto-complete this part of my translation'"



- **Productivity gain highly dependent on available <u>data-sources</u> and <u>proposal strategy</u>**
  - Optimal configurations reduce keystrokes by 30 up to 50%
  - Avoidance of typos, impact on consistency

Subsegment Level : Auto-suggest

Data sources

- **Compilation of phrase dictionaries from TM or Corpora**
  - Decide whether co-occurrence of a word in the source segment and a word in the aligned target segment is coincidence (random) or not
  - Ex. IBM Model 1-5 and subsequent proposals
- **Terminology database**
- **Placeables**
- **User defined auto-text entries**

Subsegment Level : Auto-suggest

Strategies

- **The art is to display not too many suggestions and to avoid noise (irrelevant suggestions)**
  - Otherwise lengthy browsing offsets productivity gains
- **Typical methods:**
  - Include source language to compute auto-suggest candidates (phrase dictionary / terminology database)
  - Maximize length of suggested phrases
  - Display list on minimum length of typed-in prefix and number of suggestions smaller than a threshold

Subsegment Level – Terms, Placeables, …

- **Worth own presentation…**

Control Data Sources

- **Whereas the key technology advances are in the area of subsegment reuse and statistical machine translation (SMT), the actual productivity gains relate to the ergonomics of how systems allow users to interact, control and automate the various data sources:**
  - Access, creation, chaining, weighting of TMs
  - Access to SMT pointing to specific engines
  - Compilation of phrase dictionaries on the fly

Relevant and related Standards

Topic Level	Segment Level	Subsegment Level
Exclusion from translation through markup W3C ITS	Translation Memory TMX eXchange	Auto-suggest (dictionary based auto-completions)
"Perfect Matching" utilizing bi-lingual representations XLIFF	Automated Translation XLIFF	Placeables, Terms TBX eXchange
	Auto-propagation	Concordance

Not discussed in this presentation:
Standards for routing Translation Packages from Translation Management Systems (TMS) to professional translators

Looking forward

- **Current theme for CAT tools – reviewer productivity**
  - Inclusion of track changes and commenting mechanisms in translation editor
- **Automation in the broader production chain**

www.sdl.com