LeCoNTRA
Learner corpus consisting of English-to-Dutch news translations enriched with translation process data

https://github.com/BramVanroy/LeCoNTRA

Goal
Create a dataset for translation difficulty research

- English-to-Dutch news translation
- Contains translation process data as proxy for cognitive effort
- Many translations per translator for ML experiments
- Limited number of pre-existing datasets (Daems, 2016; Vanroy, 2021)
- High quality segmentation and alignment
- Useful for translation process research, learner corpus research …

Text selection
- Selected from the Dutch Parallel Corpus (Macken et al., 2011)
- 50 texts from The Independent (EN) with reference translations from De Morgen (NL; P04)
- Different news domains (science, welfare, environment, culture …)
- Average readability score of 74.89 (<60 = hard)

Data collection
- Work from home with Translog-II to record keystrokes (Carl, 2012)
- Fixed fee and duration (125 euros, 10 hours)
- Quality over quantity: not all texts needed to be translated
- Processing with Translation Process Research Database (Carl et al., 2016)
- Manual re-segmentation of tokens and sentences (cf. Mantis)
- Manual word alignment with YAWAT (Carl et al., 2016)
- P01’s texts are most lexically rich
- P02’s texts are easiest to read
- P02’s texts are least literal in terms of word order (word cross)

Translators
- P01’s texts are most lexically rich (MSTR)
- P02’s texts are easiest to read (Douma-Flesch Reading Ease)
- P02’s texts are least literal in terms of word order (word cross)

PO2 translated a lot slower than others…

PO2 takes ~149s per sentence, almost 2x P03 and 3x P01!

- In process data, we find that on average P02 revises a segment once on average (Nedit, Nedit~2 means a segment has been revised once)
- Slower ≠ worse quality, nor vice versa!

MANTIS
Manual text segmentation. A tool for manually correcting token and sentence segmentation in the browser

https://github.com/BramVanroy/mantis

- Easy-to-use web interface
- Token and sentence segmentation of bi-texts
- Sentence alignment
- For now, focused on the Translation Process Research DataBase format
- Open source (ReactJS+FastAPI)

References


References


References
