FACULTY OF ARTS AND PHILOSOPHY

LT3, LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

Bram Vanroy, Lieve Macken

LECONTRA: A LEARNER CORPUS OF ENGLISH-TO-DUTCH NEWS TRANSLATION

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

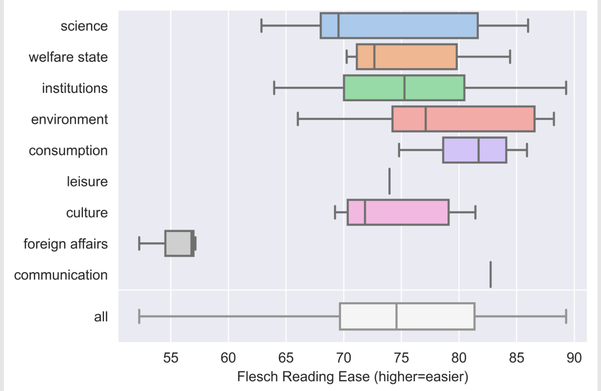
- English-to-Dutch news translation
- Contains translation process data as proxy for cognitive effort
- Many translations per translator for ML experiments
- Limited number of pre-existing datasets (Daems, 2016; Vanroy, 2021)
- High quality segmentation and alignment
- Useful for translation process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

Domain	MSTTR	Sent.	Token	FRES
foreign affairs	70.83	13.06	4.75	55.40
science	72.50	13.36	4.30	73.80
leisure	76.00	12.71	4.17	73.97
culture	74.19	14.44	4.17	74.48
welfare state	75.21	11.83	4.17	75.29
institutions	74.83	11.82	4.15	75.46
environment	75.06	12.14	4.17	78.83
consumption	75.14	12.40	4.12	81.02
communication	75.50	10.80	4.18	82.74
all	74.59	12.60	4.22	74.89

Table 1: Source text complexity variables for all domains and the whole corpus ("all"): mean segmental type-token ratio (MSTTR), average sentence (in tokens) and token lengths (in characters), and average readability scores. Sorted by the latter.



Participants

- 3 students of Master in Translation
- Dutch natives, English as language of study
- Lexical decision task scores
- TICQ metadata (Schaeffer et al., 2020)

ID	Languages	Texts	LexTALE
P01	EN-RU-NL	T01-T28;T30-T47	91.25%
P02	EN-FR-NL	T01-T24	81.25%
P03	EN-RU-NL	T01-T41	85.00%
P04		T01-T50	

Table 2: An overview of the student translators for the LeConTra corpus, the languages that they study, the texts that they translated and their LexTALE scores. P04 contains the original, published, translations as present in the DPC corpus.

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 (Germann, 2008)

Translations

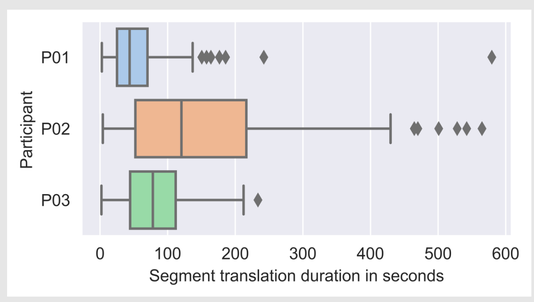
- P01's texts are most lexically rich (MSTTR)
- P02's texts are easiest to read (Douma-Flesch Reading Ease)
- P02's texts are least literal in terms of word order (word cross)

Texts	Part.	MSTTR	Sent.	Token	DFRES	word cross
1-24	P03	72.25	11.67	4.75	56.99	23.18
	P01	74.98	11.49	4.73	57.88	26.97
	P02	72.16	12.21	4.63	61.96	20.40
	P04	74.21	10.63	4.72	60.27	22.66
1-28; 30-41	P03	72.33	11.78	4.79	56.43	25.58
	P01	74.24	11.69	4.75	57.45	30.22
	P04	74.15	10.70	4.78	58.39	24.96

Table 3: Target text complexity variables of the translations of all participants. Mean segmental type-token ratio (MSTTR), average sentence (in tokens) and token lengths (in characters), average readability scores, average word order changes (cross). Sorted by readability scores per "Texts" group.

P02 translated a lot slower than others…
What can be the cause of that?

- P02 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!



Texts	Part.	Dur (s)	Nedit
1-24	P01	54.38	1.30
	P02	148.65	2.09
	P03	83.22	1.28
1-28; 30-41	P01	53.24	1.25
	P03	80.29	1.19

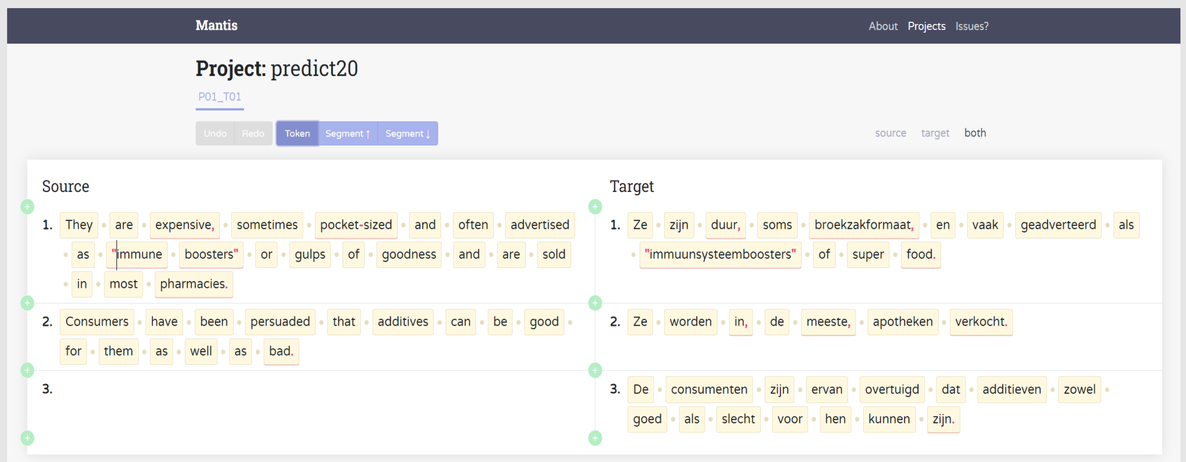
Table 4: Average translation duration per segment in seconds and average number of edits per segment.

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

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In Proceedings of the International Conference on Language Resources and Evaluation, pages 4108–4113, Istanbul, Turkey.

Carl, M., Schaeffer, M. J., and Bangalore, S. (2016). The CRITT translation process research database. In Michael Carl, et al., editors, New directions in empirical translation process research, New frontiers in translation studies, pages 13–54. Springer, Cham, Switzerland.

Daems, J. (2016). *A translation robot for each translator.* PhD thesis, Ghent University, Ghent, Belgium.

Germann, U. (2008). Yawat: Yet another word alignment tool. In *Proceedings of the ACL-08: HLT Demo Session*, pages 20–23, Columbus, Ohio, June. Association for Computational Linguistics.

Macken, L., De Clercq, O., and Paulussen, H. (2011). Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta: Journal des traducteurs*, 56(2):374–390.

Schaeffer, M., Huepe, D., Hansen-Schirra, S., Hofmann, S., Muñoz, E., Kogan, B., Herrera, E., Ibáñez, A., and García, A. M. (2020). The Translation and Interpreting Competence Questionnaire: an online tool for research on translators and interpreters. *Perspectives*, 28(1):90–108, January.

Vanroy, B. (2021). *Syntactic Difficulties in Translation.* Ph.D. thesis, Ghent University, Ghent, Belgium.

bram.vanroy@ugent.be Bram Vanroy @bramvanroy / @lt3ugent

GHENT UNIVERSITY