Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1807–1816
Marseille, 20-25 June 2022

LeConTra: A Learner Corpus of English-to-Dutch News Translation

Bram Vanroy, Lieve Macken
LT^3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
bram.vanroy@ugent.be, lieve.macken@ugent.be

Abstract

We present LeConTra, a learner corpus consisting of English-to-Dutch news translations enriched with translation process data. Three students of a Master's programme in Translation were asked to translate 50 different English journalistic texts of approximately 250 tokens each. Because we also collected translation process data in the form of keystroke logging, our dataset can be used as part of different research strands such as translation process research, learner corpus research, and corpus-based translation studies. Reference translations, without process data, are also included. The data has been manually segmented and tokenized, and manually aligned at both segment and word level, leading to a high-quality corpus with token-level process data. The data is freely accessible via the Translation Process Research DataBase and GitHub, which emphasises our commitment of distributing our dataset. The tool that was built for manual sentence segmentation and tokenization, Mantis, is also available as an open-source aid for data processing.

Keywords: translation corpus, learner corpus, empirical translation studies, keystroke logging, translation process

1. Introduction

All empirical research starts with a need for data, and depending on the research topic, this data can be hard to find. For our own research, we were specifically concerned with using behavioral process data (as measured by for instance keystroke logging and eye tracking) for translation difficulty to answer questions such as: how does the translation behavior provide insights about difficulties that a translator has encountered? Although such datasets exist, they are scarce for English-to-Dutch translation, the exception being the limited datasets of Daems (2016) and Vanroy (2021). Furthermore, to get plenty of behavioral data per subject, we were interested in a dataset with many translations per translator to cover a wide range of linguistic phenomena, rather than plenty translations per source text, so we opted to have a few translators work on many texts rather than many translators on few texts.

As a result, we collected a dataset for English-to-Dutch translation with behavioral data in the form of keystroke logging information, and we make it freely available. We are especially contributing to two fields: learner corpus research (LCR) and translation process research (TPR). On the one hand the dataset that we publish contains translations of students of Translation as well as reference translations (Sec. 3.2.1), and on the other we include translation process data in the form of keystroke logging information.

In what follows we first provide some non-exhaustive background about translation process research and translation difficulty, the field and topic in whose light this dataset was originally created, as well as existing endeavours in learner translation corpora (LTC). Then we discuss the dataset itself, with a focus on the source texts, the participants and data processing method (including a new tool for manual text segmentation). Some observations are then made considering the translations and the process data. We conclude the paper with suggestions of future work and potential use-cases for our dataset and the accompanying tool.

2. Related Research

For decades, researchers in Translation Studies have argued about whether or not an original source text can truly be translated. Structural Linguists such as Jakobson (1971) (originally published in 1959) argue that full translation equivalence between a source text and a translation are seldom encountered. In a Structuralist view where form and meaning are strongly connected to each other, this idea is similar to how two synonyms are rarely exactly the same. Nida (1964) notes that the translator's objective is to create a translation that is most similar to the source text, where "similarity" is dependent on the goal that the translation must achieve. However, "no fully exact translation" is possible (p. 156). Catford (1965) devotes a whole chapter (Ch. 14) of his *A Linguistic Theory of Translation* to "The Limits of Translatability". He discusses, among other things, how linguistic differences between the source and target languages (SL; TL) may elicit translation failure or incompatibility. Furthermore, semantic and culture differences can hinder a true transposition due to how language and culture are intertwined. Similar to Catford (1965), Baker (2011) (first published in 1992), suggests that different realisation of specific phenomena in the language systems (e.g., morphology and syntax), can hinder the translation procedure. In summary, whether or not texts can be translated entirely, including grammatical, semantic, cultural facets, depends highly on the equivalence between SL and TL. When equivalence does not exist for one of these facets or even for a specific source text unit, the translator will

be faced with a translation difficulty that needs be resolved.

As is clear from the discussion above, translation feasibility and difficulty have been a topic of discussion for a long time. And while the difficulties of translation can be investigated after the fact by means of questionnaires or self-assessment scales such as the NASA task load index (TLX) (Hart and Staveland, 1988), these approaches are highly subjective. The topic of translation difficulty has been adopted by fields such as psycholinguistics and empirical translation research, though, which provide a strong foundation of research to difficulty in experimental research. Notable, here, is the work of Campbell (1999), who embraced the cognitive aspect of translation supported by empirical studies to model how translators may come up with specific translation decisions (continued in Campbell (2000)). Equally important for this topic of translation difficulty are the methodological developments in translation process research (TPR), which is interested in the process of creating the translation and how the final text is a product of that process. Such process data can provide objective insights into the decisions and difficulties a translator encountered while translating.

Historically, think-aloud protocols (e.g., Gerloff (1986), TAP) have often been used in TPR to get a grasp of what a participant is thinking during an experiment, and by extension the difficulties that they are facing. During translation, they are asked to utter what they are thinking and doing, clarifying the decisions that they make or the points where they struggle. However, such a conscious process intervenes considerably with the translation process and may distract the participant from the task at hand (Krings, 2001), leading to side effects in completing it, such as slower processing or over-awareness of their own work. For this reason, Carl et al. (2008) proposed to employ objective user activity data as an approximation of what is going on in the mind of a translator. This data is recorded during translation by means of, for instance, keystroke logging software or eye-tracking equipment. In the words of Leijten and Van Waes (2013), the "main rationale behind keystroke logging is that writing fluency and flow reveal traces of the underlying cognitive processes" (p. 360). It is for instance possible to dive into the translation duration of single tokens or whole texts, the number of revisions or typing mistakes (typing efficiency), or the pause behavior of a translator (O'Brien, 2006; Kumpulainen, 2015; Lacruz et al., 2012). Eye tracking, while not used in the current corpus, is a powerful research method as well that can measure the eye fixations and gaze durations of participants (Carl et al., 2010; Daems, 2016; Jakobsen, 2011; Schaeffer et al., 2016). In sum, behavioral data has become of high value for translation process research in general and translation difficulty research specifically.

In addition to translation process research, our dataset also aims to add value to learner corpus research. As will be explained in the following sections, the translations presented in the dataset are created by Dutch-native advanced learners of English. Learner translation corpora are rather scarce but some projects have yielded impressive datasets, with recent initiatives showing much promise for the field.

The Russian Learner Translator Corpus (RusLTC) contains bidirectional English-Russian translations that were collected from 2014 onward (Kutuzov and Kunilovskaya, 2014). The language learners were Russian students from several Russian universities, and the collected data were taken from assignments, exams, and contests. The data has also been quality-annotated, which allowed for the work of Kunilovskaya and Lapshinova-Koltunski (2019). Those authors used RusLTC for research on translationese, the difference between original text and translated text. They found that there is no direct correlation between the level of translationese in a translation and its quality in the RusLTC corpus.

Another learner translation corpus is the LTC-UPF corpus of the University of Pompeu Fabra in Barcelona (Espunya, 2014)f. It contains English-Catalan written translations of students of the Translation and Interpreting programme. Every translation has been sentence-aligned and enriched with automatic linguistic annotations. Furthermore, the translations were error annotated by language instructors. The data is searchable by means of a query tool so that instructors can search by academic year, course, text type, and so on. The corpus was specifically developed as "an aid for teachers" (p. 39), who can gather insights from the translations and apply it to designing an appropriate curriculum for their students. Additionally, the authors suggest that it may be a valuable resource for students, although it is recommended that they only make use of the corpus with supervision as to not over-emphasize best vs. worst performances and errors.

A final learner corpus initiative that we would like to highlight is the Multilingual Student Translation (MUST) (Granger and Lefer, 2020). It is a fairly recent initiative that spans many institutions and highlights the importance and possibilities of international collaboration. The dynamic corpus can be used by collaborating partners within the MUST network, who in turn are encouraged to add to the existing source texts by providing new student translations. The corpus can also be expanded with new texts and translations as long as an extensive set of metadata is supplied. It should include metadata regarding the source texts, the translation tasks, and the translator students. Similar to LTC-UPF above, the founders of MUST suggest that the corpus can be used to improve the pedagogy of teachers. Moreover, it opens doors for data-driven learning, where students are faced with a faulty translation with error-type annotation rather than a source text to better understand translation errors. The authors also explicitly encourage the use of the corpus for research,

for instance for variation and choice analysis, general corpus linguistics, research involving the translators' metadata, and so on.

With LeConTra, we provide a hybrid corpus compared to the translation process corpora and LTC. On the one hand, our data is enriched with translation process data, manually segmented on the word and sentence level, and manually aligned. On the other hand its translations are collected from advanced learners of English with plenty of metadata. As such, and similar to the corpora above, we aim to contribute to different types of research such as translation process research, learner corpus research, and corpus-based translation studies.

3. LeConTra

We present the LeConTra dataset (**Le**arner **Co**rpus of English-to-Dutch **N**ews **Tra**nslation). The source texts, its student translations, and professional reference translations are provided as a public data collection study on the Translation Process Research DataBase ((Carl et al., 2016), TPR-DB), a web interface that computes useful metrics for the recorded process data (if any), such as token-level typing duration, number of revisions per segment, and so on. A stable, final version of the data is also available on GitHub, which also includes the collected metadata (cf. below).[1] Translation process data in the form of keystroke logging information is included for the student translations.[2]

3.1. Source Text

3.1.1. Selection

LeConTra contains student translations of source texts that were selected from the Dutch Parallel Corpus (Macken et al., 2011, DPC). The source texts in DPC were originally collected in 2010-2011. The subset used in LeConTra was selected from the part of the corpus that contains English source texts published in the newspaper The Independent[3] and that were professionally translated and published in the Belgian newspaper De Morgen[4] in Dutch. That means that in addition to the learner translations, we also include the professional, quality-approved and published, translations as references. It should be emphasized that it is unlikely that these reference translations were all made by the same translator - they merely serve as *a* reference translation.

Texts were first selected based on their difficulty as perceived by the first author of this paper, who has a formal background in English and Dutch linguistics. Special attention was paid to terminology, sentence length, and complexity (e.g., the level of co-reference, abstractness, syntactic constructions). The goal was to select texts that were not too easy to translate but that did not require external resources (dictionaries, knowledge bases) to produce a good translation. As will be discussed in Section 3.2.1, students were not allowed to make use of external resources. Second, the text structure was taken into account, particularly because this corpus was created in light of translation process research, and the format of the texts needed to accommodate the tools that were to be used by the student translators. They made use of the program Translog-II (Carl, 2012) to record their translation process (keystrokes), as will be discussed below. As such, each text was limited in size to fit the interface of this tool with an average text length of 10 sentences (between six and 16 sents.). To reach this limited size, the original DPC source texts were manually trimmed or split across separate source texts into meaningful chunks of the desired length. The professional translations that we include have been trimmed accordingly to correspond to the source texts.

Table 5 in the appendix contains a full overview of all individual texts, including their news-related subdomain (as taken from DPC metadata), the number of source tokens that the text contains after trimming, and the number of sentences.

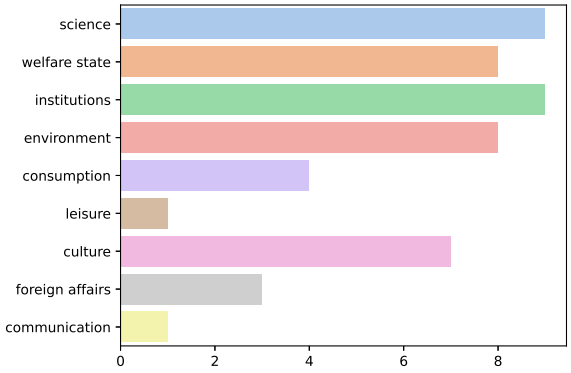

Figure 1: Bar plots indicating the number of texts per journalistic domain.

In Figure 1, the distribution of the different subdomains of the texts are given. Not all domains are equally present, which is important to consider for further statistics. Especially the communication (1), leisure (1), and foreign affairs (3) domains are less represented, so statistical information concerning these should be taken as tendencies.

3.1.2. Statistics

To gauge the text complexity of the source texts, we present mean segmental type-token ratio (MSTTR), average sentence and token lengths, and average Flesch Reading Ease Scores in Table 1. In this section we

[1] https://github.com/BramVanroy/LeConTra

[2] The following URL provides details on how to access public datasets in the translation process research database: https://sites.google.com/site/centretranslationinnovation/TPR-DB#h.p_bx0xkGDxfVcs

[3] https://www.independent.co.uk/

[4] https://www.demorgen.be/

break down the statistics per sub-domain but also provide details for the whole corpus.

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.

Type-token ratio quantifies the lexical diversity (or richness) in a corpus and therefore measures text complexity based on vocabulary. However, this is dependent on corpus length, which are not comparable across our domains. A variation of TTR, mean segmental TTR (Johnson, 1944), takes that into account. It divides the corpus into segments of equal length (100 tokens in our case), and calculates TTR for each separate segment and then averages all the scores. In Table 1, MSTTR for all the different domains is given (multiplied by 100). The closer that this value is to 100, the more diverse (or rich) the vocabulary use is. The science and foreign affairs domains stand out with a markedly lower lexical diversity score than the other domains.

Readability measures are another way to approximate text complexity (Daems, 2016; Sharmin et al., 2008). They focus on surface characteristics and are calculated based on factors such as sentence length and number of syllables. We provide Flesch Reading Ease Score (Flesch, 1949, FRES), a readability score between 0-100, for which a score of 60 and lower is considered difficult. We find that, on average, texts have a FRES of 74.89 (M = 74.57, SD = 8.34), the easiest text has a score of 89.28 and the hardest 52.26. On average the selected texts are therefore not very difficult. A reasonable reading difficulty level may have been a deliberate quality criterion of the source text provider, The Independent, to make the news articles accessible to a broad audience.

In terms of sub-domains, it is notable that especially the texts of news concerning foreign affairs are hard to read (Fig. 2), although it should be noted that there is only a very limited number of texts per domain, and in the case of foreign affairs all three selected texts are part of the same original DPC text.

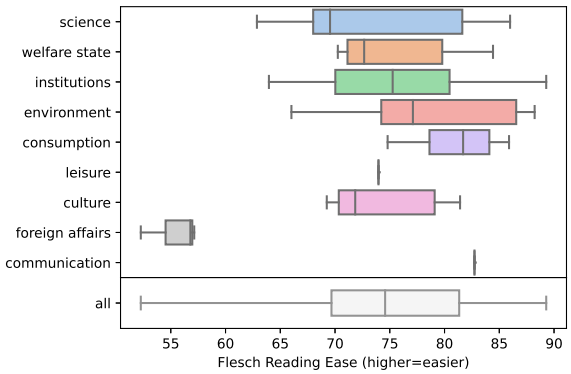

Figure 2: Box plots indicating the readability of source texts per journalistic domain.

3.2. Methodology

3.2.1. Participants

The source texts were translated by three students of the Master's programme in Translation at Ghent University at the department of Translation, Interpreting, and Communication. These translators were native Dutch speakers and had picked at least Dutch and English as translation languages in their programme. All students were female but it is generally assumed that gender of participants is not a decisive factor in cognitive-related research (Hvelplund, 2011). Metadata was collected after the task was completed with TICQ, the Translation and Interpreting Competence Questionnaire (Schaeffer et al., 2020), which we pseudonymized and include in our GitHub repository. Students were given a fixed fee of 125 euros for 10 hours of employment and were asked to translate as many of the provided texts as they could within the given time frame as well as they could. It was emphasized that the quality of the translation had precedence over the quantity that they could produce. Therefore, although a selection of 50 texts were made, it is clear from Table 2 that not all participants translated the same number of texts, i.e., that they translated at different speeds.

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.

To probe the English vocabulary knowledge of the participants, they were asked to complete an unspeeded lexical decisions task, specifically the English LexTALE test ((Lemhöfer and Broersma, 2012), Lexical

Test for Advanced Learners of English).[5] The creators of the evaluation tool report that a large group of Dutch and Korean advanced learners of English achieved an average score of 70.7, so our students have a very good understanding of the English lexicon.

As mentioned, translation process data in the form of keystroke logging was collected with the computer program Translog-II (Carl, 2012). The program provides a split-screen interface, displaying the source text to translate on one side and providing an input field on the other where the translation can be typed. While typing in that field, a translator's keystrokes are recorded. Because no additional equipment is necessary to run this software, participants were able to work from home on a Windows computer, which was the preferred method of working due to the COVID-19 pandemic. Students were not allowed to make use of any external resources (books, dictionaries, knowledge bases, and so on) as to ensure that the recorded process data (such as duration and keystrokes) was limited to the translation process itself and not distorted by accessing other resources.

3.2.2. Data Processing

After the translators finished their work, the Translog-II output files were uploaded to the TPR-DB (Carl et al., 2016). The TPR-DB automatically tokenizes and sentence segments the data to calculate metrics for different linguistic units. Tokenization is the process of splitting characters in meaningful parts (typically visualized by adding a space between parts), e.g., *you're* → *you 're*. Sentence alignment is not done automatically, but source and target sentences are simply ordered sequentially, assuming a one-to-one correspondence of sentences. Although both of these automatic segmentation processes are useful, they are also prone to errors. The first author of this paper therefore manually corrected the tokenization, sentence segmentation, and sentence alignment between the source text and translations. The additional scripts that were developed to aid such manual endeavours, have been improved and incorporated into an interface that we call Mantis (Sec. 3.2.3, Manual Text Segmentation). It provides a user-interface for researchers to manually segment their data on the token and sentence level and is a by-product of this data collection study.

After this manual process of segmentation, a student-annotator was hired to align all the source texts to their translations on the word level.[6] Word alignment is the process of linking a source word to its (linguistically) corresponding target unit. Such alignments allow the TPR-DB to calculate bilingual metrics that incorporate both the source and target text, e.g., concerning word order changes or involving the duration to produce the target word of a specific source token. The annotator worked from home and made use of the YAWAT interface for word alignment (Germann, 2008).

Finally, after all preprocessing steps were completed, the TPR-DB calculated meaningful metrics pertaining to the process data, e.g., pauses, duration, editing behavior, and the final translation, e.g., entropy metrics to investigate the likelihood of a translation. These metrics are then stored as spreadsheet tables that can be downloaded and analyzed.

3.2.3. Mantis

While it is evident that the TPR-DB is a useful tool for TPR, we found that automatic tokenization and sentence segmentation is prone to errors. Because TPR is often interested in specific tokens at a fine-grained level as well as in an accurate correspondence between source and target segments, it is of paramount importance that segmentation is of outstanding quality. Therefore, we developed an interface to manually correct the tokenization and sentence segmentation of the TPR-DB, called Mantis (Manual Text Segmentation).[7] For now, the tool should be installed by users locally on their own device or server.

After installation, users should download the proposed data segmentation of the TPR-DB by downloading the "alignments" from that interface. Then, they can easily start the tool on their own device and upload the alignment data. Per project, per participant, and per text, the user can then verify and change the tokenization of sentences, as well as their sentence segmentation and alignment. All of this can be done by simply using the mouse. When a user has verified all text segmentation, they can save and download their changes, and upload the updated files to the TPR-DB, which in turn can then calculate measures for the manually, corrected linguistic units.

In Figure 3, a screenshot of the tool's interface is given. By default, the tool is set to tokenization (for word-level segmentation) rather than Segment (for sentence-level segmentation). The screenshot shows that the user has their cursor between *"* and *immu...*. If they were to click in this position, *"* would be separated and considered a separate token. Tokens that contain punctuation are automatically underlined in red because these are often, but not always, problematic cases that have been incorrectly tokenized. A user can "detokenize" a token, i.e. glue it back together with a previous chunk, by right-clicking.

With the toolbar, users can undo and redo actions, but they can also switch to "sentence segmentation" (Segment). The arrows indicate the direction of segmentation: by clicking between two tokens, a segment is split into two and the direction decides whether the first part moves up or the second chunk moves down. If used

[5] `http://www.lextale.com/`

[6] Since our data was collected, the TPR-DB has now included SimAlign (Jalili Sabet et al., 2020) to automatically word align the data. As before, this is useful but whenever possible one may prefer manual verification of those suggestions

[7] `http://github.com/BramVanroy/mantis` Design and functionality are subject to change.

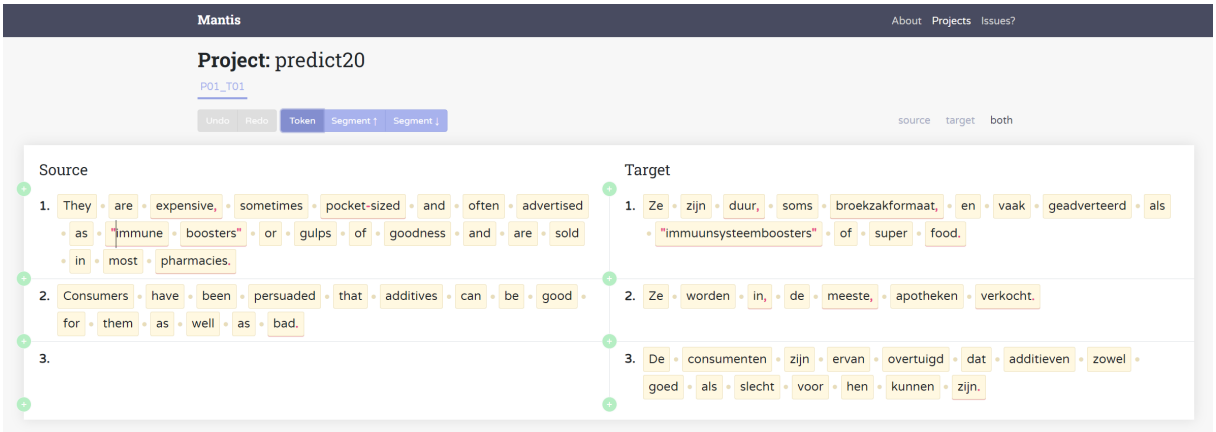

Figure 3: The interface of Mantis.

at the front of a segment with direction "down", or at the end of a segment with the direction "up", the whole segment is relocated. Users can easily add new empty rows by pressing the green "+" button.

Figure 4: Mantis example of segmentation.

This implication of segmentation in this way, is that it also enables segment alignment: one can reposition the segments of both the source and the target text until these source and target segments correspond. So in this case, we see that the translator translated the first sentence as two separate sentences. To make sure that the segments are correctly aligned, we can therefore choose to either split the source sentence in two segments or join the second target segment with the first. In Figure 4, the users opts to take the latter approach: the cursor is positioned after "verkocht." and the tool is set to "Segment up". Upon clicking, the segment "Ze worden in, de meeste, apotheken verkocht." will move up to join the first segment. The third translation segment ("De consumenten ...") will move up automatically so that all source segments are correctly aligned.

3.3. Target Texts

Because not all students translated the same number of texts, this section will only focus on two groups of texts. On the one hand those texts that were translated by all students (T01-T24), and additionally the data that was translated by P01 and P03 (T01-T28; T30-T41). The professional, published translations are included as P04, but as noted before these are likely not all written by the same translator and serve as a potential professional reference translation rather than as a single professional translator. Rather than emphasizing different domains, the focus lies on differences between translators.

3.3.1. Product Data

Similar to Section 3.1.2, we provide corpus statistics of the final products, i.e., the translations themselves that were collected. First, we include MSTTR scores as a probe for lexical richness in Table 3. Lexical richness of the translation is often used in language learning and language acquisition research (with varying results (Thomas, 2005)), and in style analyses of translators (Huang, 2015). In our dataset, we find that, perhaps coincidentally, the lexical richness of translators correspond to their *English* LexTALE scores (Table 2). As it stands for the first 24 texts, P01 shows the richest vocabulary in their Dutch translations closely followed by the reference translations. P03's translations are barely more diverse than those of P02. The trend continues in the larger subset, where P01's translations are more lexically rich than P04 and then P03.

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.

In terms of readability and limited to the first 24 texts, P02's translations are easier to read, followed by the reference translations. The translations of P03 are hardest to read on average. The readability scores for the

different participants are visualized in Figure 5, based on the Flesch Reading Ease Scores but adapted for Dutch (Douma, 1960). When taking into consideration the first 41 texts (excl. T29) translated by P01, P03 and P04, the same observation can be made: the reference translations are easiest to read, followed by P01 and then P03.

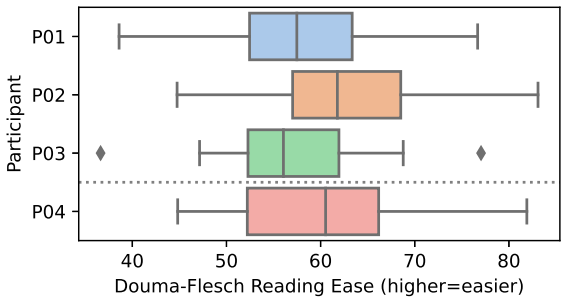

Figure 5: Box plots indicating the readability of translations per translator for the first 24 texts.

Finally, as an example of how the dataset can be used to analyze cross-linguistic divergencies, we can look at the tendencies with respect to literal, word-for-word translation. Research suggests that literal translation is the "default rendering procedure" (Tirkkonen-Condit (2005), p. 407-408), and empirical research confirms that it is often the first starting point of a translation (Carl and Dragsted, 2012). It has been found that in general novice translators hold on to this first, literal translation more strongly than professional translators (Chesterman, 2011; Englund Dimitrova, 2005). To look into syntactic literality, we can compare the average word reordering of the learner translators and the reference translations by means of the word cross metric of Vanroy et al. (2019), which quantifies how often word alignment links cross each other in a sentence translation. Put differently, how literally do translators convey the word order of the source text into their translation?

From our data ("word cross" in Table 3), we cannot directly confirm that the student translators used a more syntactically literal approach than the professional translations. In both subsets of texts it is obvious that P01 shows a tendency to translate more freely (less literal, higher word cross). On average they reorder the words a translation compared to the source text more than the other translators, even more so than the reference translation. In the first 24 texts, P02 exhibits the most literal translation behavior.

3.3.2. Process Data

One of the contributions of this dataset is that we include process data for English-to-Dutch news translation. Because the data has been processed with the TPR-DB (Carl et al., 2016), we have access to a variety of keystroke logging metrics.[8]

To illustrate this, we can verify the information of Table 2, where it was clear that P02 translated less texts than the other students by looking at translation duration information derived from the keystrokes. Indeed, in Table 4 and Figure 6, we see that for the first 24 translations, P02 spend considerably more time on their work with an average translation time of $148.65s$ per segment. This is almost three times the average translation time of P01 ($54.38s$). P03 is around $30s$ slower than P01 for both text subsets.

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.

Note that this analysis does not include quality verification: although students were asked to prioritise translation quality, their respective translation quality has not been verified. Therefore, the assumptions must not be made that the faster translator produced the best translation, or, conversely, that the slower translator was more accurate. However, the process data enables us to find clues related to the slower process of P02.

To demonstrate, the "Nedit" variable that the TPR-DB calculates is relevant (number of edits or revisions). It quantifies how often a translator worked on the translation of a segment. So if they translate the source segment in one go and move on to never return to work on the segment (the default), then that is a value of 1. But if they go to another segment and then go back to revise and change the former segment, then that counts as an additional edit. The average number of edits/revisions per segment in Table 4 explains, perhaps only in part, why P02 took so much more time to translate each segment from a process perspective. P01 and P03 have a mean of only around 1.3 edits, which means that they quite often translated the segment, and when they were satisfied moved on to another segment without ever returning to change the former segment. P02, however, has a mean of more than 2 edits, which clarifies their revision process of going back and changing a previous translation that they had already made. On average they revised their segment at least once after an initial translation. So, P02 seems to exhibit a high level of self-revision. In the larger subset of the data, P01 also revises slightly more than P03.

[8] See `https://sites.google.com/site/centretranslationinnovation/TPR-DB/features` for an overview.

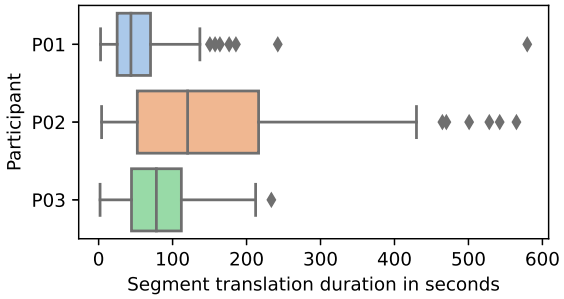

Figure 6: Box plots indicating the segment translation duration in seconds per student translator for the first 24 texts.

Although these differences in translation method should not be equated with translation quality, it is clear that translation process data can provide insight in how translators have different translation strategies. In this case, P02 revised more frequently, leading to a slower translation process, whereas the other participants first completely finished a segment and then moved on more often. Interesting, again, is that P02, who revised often, was also the participant with less knowledge of the English lexicon (Table 2). This may be indicative of a higher uncertainty about their translations, or it simply illustrates that they thoroughly self-monitor their translations. A thorough quality analysis of the translation would be necessary to provide a conclusive explanation but at least the process causing a longer translation time, namely revision, is clear.

4. Conclusion and future work

In this paper we have introduced a dataset of learner English-to-Dutch translations of news texts (LeConTra). Professional reference translations are provided alongside learner translations of students of a Master's programme in Translation. For the student translations, translation process data is included as keystroke logging. The data was manually tokenized, sentence segmented, and sentence aligned. Participants' metadata was pseudonymized and is also included.

We provided text statistics concerning text complexity (readability and type-token ratio) for both the source texts and the translations, alongside average word and sentence lengths. Having access to reference translations allow comparisons to be made with the learner translations. As an example, we compared the translation literality as operationalized by a word reordering metric, but found little discernable differences between professionals and students.

Complementary to the product data, we illustrated the usefulness of translation process data by looking into the differences between translation duration of the student translators in Section 3.3. It was clear that not all translators had completed the same number of texts in the same time span, receiving the same compensation. We found that the process data indicated that their production duration differed, at least in part, because they revised their translations in distinct ways. One student, P02, revised and edited more extensively than the others, leading to longer translation times.

In addition to the dataset itself, we also make Mantis available as an open-source tool for manual text segmentation. It was developed as a support aid to deliver high quality preprocessing for our data. In addition to its current functionalities of tokenization, segmentation, and segment alignment, also word alignment can be incorporated in the future. It could therefore become a modern replacement for YAWAT (Germann, 2008) with "batteries included", i.e., including the option to manually segment sentences and tokenize words before starting segment and word alignment (YAWAT focuses on word alignment). Currently, Mantis is focused on being used in tandem with the TPR-DB. We aim to opening it up to be used with a variety of formats that are common in the use of parallel corpora (e.g., plain text, vertical format, horizontal format with separator, TMX).

With LeCoNTra, we specifically created a dataset with a focus on a lot of translations per participant. This is relevant to our research on translation difficulty where we were in need of much behavioral data per translator profile to uncover a variety of linguistic phenomena. As such, the analyses that were presented in this paper are limited in terms of generalisation across participants, but useful in terms of comparisons between individuals. We hope that by adding our data to the TPR-DB (Carl et al., 2016), which encourages collaboration and shared insights, that other researchers will make use of our selected source texts, for instance contributing an English-to-French or English-to-Japanese version of LeConTra. Such additions would allow us, and others in the field, to not only analyse our own translations but additionally, and excitingly, extent research to comparative studies across translations of the same source texts in different target languages. We are interested in using this dataset and its process data and other contributions with different target languages for research on translation difficulty, but in addition to TPR, and facilitated by the inclusion of professional reference translations, the corpus can be used for a corpus-based approach to language learning research as well.

5. Bibliographical References

Baker, M. (2011). *In other words: A coursebook on translation*. Routledge, Abingdon, UK, 2 edition.

Campbell, S. (1999). A cognitive approach to source text difficulty in translation. *Target*, 11(1):33–63.

Campbell, S. (2000). Choice network analysis in translation research. In Maeve Olohan, editor, *Intercultural faultlines: Research models in translation studies*, pages 29–42. St. Jerome, Manchester, UK.

Carl, M. and Dragsted, B. (2012). Inside the monitor model: Processes of default and challenged trans-

lation production. *Translation: Computation, Corpora, Cognition*, 2(1):127–145.

Carl, M., Jakobsen, A. L., and Jensen, K. T. H. (2008). Studying human translation behavior with user-activity data. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, pages 114–123, Barcelona, Spain.

Carl, M., Kay, M., and Jensen, K. T. H. (2010). Long distance revisions in drafting and post-editing. In *Proceedings of CICLing 2010*, pages 193–204, Iaşi, Romania.

Catford, J. C. (1965). *A linguistic theory of translation: An essay in applied linguistics*. Oxford University Press.

Chesterman, A. (2011). Reflections on the literal translation hypothesis. In Cecilia Alvstad, et al., editors, *Methods and strategies of process research integrative approaches to translation studies*, volume 94, pages 23–35. John Benjamins Publishing Company, Amsterdam ; Philadelphia.

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

Douma, W. H. (1960). De leesbaarheid van landbouwbladen : een onderzoek naar en een toepassing van leesbaarheidsformules. *Bulletin*, 17:54.

Englund Dimitrova, B. (2005). *Expertise and explicitation in the translation process*, volume 64. John Benjamins Publishing Company, Amsterdam ; Philadelphia.

Flesch, R. (1949). A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221.

Gerloff, P., (1986). *Second Language Learners Reports on the Interpretive Process: Talk-aloud Protocols of Translation*, pages 243–262. Gunter Narr Verlag, Tubingen.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In Peter A. Hancock et al., editors, *Advances in psychology*, volume 52 of *Human Mental Workload*, pages 139–183. North-Holland, January.

Huang, L. (2015). *Style in Translation: A Corpus-Based Perspective*. New Frontiers in Translation Studies. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. PhD thesis, Copenhagen Business School, Frederiksberg.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In Cecillia Alvstad, et al., editors, *Methods and strategies of process research: Integrative approaches in translation studies*, volume 94 of *Benjamins Translation Library*, pages 37–55. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Jakobson, R., (1971). *On linguistic aspects of translation*, volume 2, pages 260–266. De Gruyter Mouton.

Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November. Association for Computational Linguistics.

Johnson, W. (1944). I. A Program of Research. *Psychological Monographs*, 56(2):1–15.

Krings, H. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. The Kent State University Press, Kent, Ohio, US.

Kumpulainen, M. (2015). On the operationalisation of 'pauses' in translation process research. *Translation & Interpreting*, 7(1):47–58.

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2019). Translationese Features as Indicators of Quality in English-Russian Human Translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56, Varna, Bulgaria, September. Incoma Ltd., Shoumen, Bulgaria.

Lacruz, I., Shreve, G. M., and Angelone, E. (2012). Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In *Proceedings of AMTA 2012 Workshop on Post-Editing Technology and Practice*, pages 21–30, San Diego, California, USA.

Leijten, M. and Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Nida, E. (1964). *Toward a science of translating*. E.J. Brill, Leiden, Netherlands.

O'Brien, S. (2006). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures*, 7(1):1–21.

Schaeffer, M., Carl, M., Lacruz, I., and Aizawa, A. (2016). Measuring cognitive translation effort with activity units. *Baltic Journal of Modern Computing*, 4(2):331–345.

Sharmin, S., Špakov, O., Räihä, K.-J., and Jakobsen, A. L. (2008). Effects of time pressure and text complexity on translators' fixations. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, pages 123–126, New York, NY, USA, March. Association for Computing Machinery.

Thomas, D. (2005). *Type-token Ratios in One Teacher's Classroom Talk: An Investigation of Lexical Complexity*. Ph.D. thesis, University of Birmingham, United Kingdom.

Tirkkonen-Condit, S. (2005). The monitor model revisited: Evidence from process research. *Meta: Journal des Traducteurs/Meta: Translators' Journal*, 50(2):405–414.

Vanroy, B., Tezcan, A., and Macken, L. (2019). Predicting syntactic equivalence between source and target sentences. *Computational Linguistics in the Netherlands Journal*, pages 101–116.

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

6. Language Resource References

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.

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.

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation*, 48(1):33–43, March.

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.

Granger, S. and Lefer, M.-A. (2020). The Multilingual Student Translation Corpus: A Resource for Translation Teaching and Research. *Language Resources and Evaluation*, 54(4):1183–1199, December.

Kutuzov, A. and Kunilovskaya, M. (2014). Russian Learner Translator Corpus. In Petr Sojka, et al., editors, *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 315–323, Cham. Springer International Publishing.

Lemhöfer, K. and Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44(2):325–343, June.

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.

A. Overview source texts

ID	DPC ID	Domain	Tokens	Sent.	Transl.
T01	dpc-ind-001632	science	279	10	3
T02	dpc-ind-001630	welfare state	127	6	3
T03	dpc-ind-001630	welfare state	220	9	3
T04	dpc-ind-001633	institutions	157	8	3
T05	dpc-ind-001639	environment	318	10	3
T06	dpc-ind-001641	welfare state	271	11	3
T07	dpc-ind-001642	consumption	224	9	3
T08	dpc-ind-001642	consumption	209	8	3
T09	dpc-ind-001642	consumption	234	10	3
T10	dpc-ind-001644	institutions	225	12	3
T11	dpc-ind-001644	institutions	181	13	3
T12	dpc-ind-001644	institutions	190	11	3
T13	dpc-ind-001648	institutions	235	10	3
T14	dpc-ind-001648	institutions	326	12	3
T15	dpc-ind-001648	institutions	331	13	3
T16	dpc-ind-001651	environment	214	10	3
T17	dpc-ind-001652	leisure	267	11	3
T18	dpc-ind-001657	science	247	10	3
T19	dpc-ind-001657	science	283	11	3
T20	dpc-ind-001658	culture	268	9	3
T21	dpc-ind-001658	culture	319	12	3
T22	dpc-ind-001659	environment	303	9	3
T23	dpc-ind-001716	science	223	10	3
T24	dpc-ind-001716	science	219	9	3
T25	dpc-ind-001718	culture	298	10	2
T26	dpc-ind-001720	foreign affairs	293	13	2
T27	dpc-ind-001720	foreign affairs	224	9	2
T28	dpc-ind-001720	foreign affairs	175	6	2
T29	dpc-ind-001721	culture	313	13	1
T30	dpc-ind-001721	culture	294	11	2
T31	dpc-ind-001723	culture	192	7	2
T32	dpc-ind-001723	culture	251	9	2
T33	dpc-ind-001724	institutions	255	9	2
T34	dpc-ind-001724	institutions	322	11	2
T35	dpc-ind-001725	science	268	12	2
T36	dpc-ind-001725	science	253	10	2
T37	dpc-ind-001725	science	314	11	2
T38	dpc-ind-001728	welfare state	312	11	2
T39	dpc-ind-001729	welfare state	238	12	2
T40	dpc-ind-001729	welfare state	215	11	2
T41	dpc-ind-001734	communication	270	13	2
T42	dpc-ind-001736	environment	317	9	1
T43	dpc-ind-001737	consumption	176	9	1
T44	dpc-ind-001740	welfare state	200	10	1
T45	dpc-ind-001740	welfare state	191	9	1
T46	dpc-ind-001743	science	278	10	1
T47	dpc-ind-001746	environment	223	11	1
T48	dpc-ind-001746	environment	271	16	1
T49	dpc-ind-001746	environment	243	16	1
T50	dpc-ind-001752	environment	235	11	1
Total			12,491	522	

Table 5: The identifiers of LeConTra texts, the DPC IDs where they have been taken from, their news sub-domains, and the number of source words and sentences in the final source texts. The last column indicate how many students translated the text.