Distinguishing translations from non-translations and identifying (in)direct translations’ source languages[1]

Laura Ivaska
University of Turku

Abstract

The scope of this study is threefold. First, machine learning will be applied to distinguish translated from non-translated Finnish texts. Then, it will attempt to identify the source languages of the translated Finnish texts. Finally, the source language identification will be tested with indirect translations, that is, with translations made from translations. The three underlying research questions are: 1) Can translated Finnish be distinguished from non-translated Finnish? 2) Can the source languages of Finnish translations be identified? 3) If the answer to question 2 is yes, then what happens when the method is applied to indirect translations; will the analysis identify the ultimate source language, the mediating language, or neither?

This study is based on the hypothesis that translated language contains traces of the source language (Toury 1995). The corpus of the study consists of non-translated Finnish prose, Finnish prose literature translations made from English, German, French, Modern Greek, and Swedish, as well as indirect translations from Modern Greek into Finnish via English, German, French, and Swedish. The analyses are based on cluster analysis and support vector machines using the frequencies of the most frequent lemmatized words.

Results show that translated and non-translated Finnish can be distinguished by using machine learning techniques. Support vector machine-based source language identification, however, was only partially successful, while a cluster analysis suggested that there is coherence within a group of texts translated from the same source language and variation between the groups of texts with different source languages. Clustering was further tested with indirect translations, and the results were mixed: six of the thirteen tested indirect translations clustered with

[1] Thank you to Ilmari Ivaska, Jaakko Kankaanpää, Einari Aaltonen, Jonna Joskitt-Pöyry, Arja Kantele, Outi Menna, Jaana Nikula, Raija Rintamäki, Tähti Schmidt, Sirkka-Liisa Sjöblom, Oili Suominen, and Marja Wich. This research was funded by the Kone Foundation.

direct translations from the ultimate source language, two with translations from their mediating languages, and five with neither.

1 Theoretical background

Baker's (1993, 243) suggestion to explore translation universals, or "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" gave rise to corpus-based translation studies in the 1990s. Researchers have since studied what features translations share and what distinguishes translated language from non-translated language (e.g., Mauranen 2004; Baroni and Bernardini 2006). Also, following Toury's (1995) suggestion that translated language contains traces of the source language (SL)—a phenomenon he calls interference—other studies have focused on identifying the SLs of translations and the linguistic features that make the identification possible (e.g., Koppel and Ordan 2011; Lynch and Vogel 2012; Islam and Hoenen 2013).

The main focus of this study, however, is on whether corpus methods can identify (the SLs of) indirect translations (ITrs). In previous research on SL identification, ITrs were not commonly taken into account, which may have led to less accurate findings than what might have otherwise been obtained. Nevertheless, ITrs are interesting because they are the result of a chain of several texts/languages: the ultimate source text/language → mediating text/language → ultimate target text/language (cf. Assis Rosa, Pieta and Maia 2017), for example Greek → French → Finnish. This raises the question of if direct translations contain traces of their SLs, do ITrs contain traces of the ultimate SL, the mediating language, or both?

Recently, Ustaszewski (2018, 173) tackled the question, "Is there an effect of the pivot language on target texts in indirect translation, and is this effect strong enough to discriminate between direct and indirect translations?" However, the results of his study cannot be confirmed because the Europarl corpus that he used lacked metadata on the (in)directness of the translations. In the current study, the (in)directness of the translations and their SLs are known and, therefore, the outcome of the SL identification can be confirmed. The results of this study can, then, be applied to and facilitate further research on ITr: if the SL identification methods detect the mediating languages of ITrs, the methods could be used to uncover ITrs, an arduous task when using the current methods (cf. Ivaska 2018).

2 Materials

The corpus in this study contains different variants of Finnish: 1) non-translated prose literature (Fi–Fi); 2) literary translations from English, French, German, Modern Greek, and Swedish (En–Fi, Fr–Fi, De–Fi, Gr–Fi, and Sv–Fi, respectively); and 3) indirect translations (ITr) of Modern Greek literature translated via English, French, German, and Swedish. The texts included in the corpus are novels and, because direct translations from Modern Greek into Finnish are scarce, there is also one collection of Gr–Fi short stories in the corpus (for the sake of clarity, the latter is considered one text even though it contains texts by several authors and various translators).

The majority of the texts used in this study come from two corpora, the Corpus of Translated Finnish (CTF) (Mauranen 2004) and Intercorp (Cermak and Rosen 2014) (Table 1). Since these two corpora contain only a few translations from languages other than English, further texts were solicited directly from translators. The translations from Modern Greek were scanned and processed into an electronic text format using Adobe Acrobat Pro DC's Optical Character Recognition (OCR). The results of the OCR have not been cleaned, and thus the translations from Greek are likely to contain errors; however, since all of the Gr–Fi and ITr texts went through a similar process, the effect of the eventual errors can be expected to even out.

The texts, except for the ITrs, were divided into subcorpora according to the language variant (De–Fi, En–Fi, Fr–Fi, Fi–Fi, Gr–Fi, Sv–Fi) they represented. Then, these subcorpora were further divided into training and test subcorpora (70% and 30% of the texts, respectively; see Table 2), and, to fade out authorial/translatorial style, texts by one author or translations by one translator in a particular language pair were always included in the same subcorpus (e.g., two novels by J.K. Rowling or three En–Fi translations by Kalevi Nyytajä are all either in training or test subcorpus).

As for the 13 ITrs, their indirectness, as well as their (assumed) SLs, had already been established in an ongoing research project (Ivaska 2016; Ivaska and Paloposki 2018). Some of the ITrs are compilative, meaning that they have been made with the help of support translations, where the translator had more than one language variant of the work (or, several source texts in different languages) at their disposal while composing the translation (cf. Dollerup 2000). However, in this current study, only the primary mediating language of each translation was considered, as the role of the supporting translations was assumed to be marginal (cf. Ivaska forthcoming).

The texts were lemmatized with UDPipe (Straka and Strakova 2017, 88). Then, as is customary in the field, all the texts in each of the training and test subcorpora were shuffled at the sentence level to fade out features other than those attributable to the SL (e.g., author style; cf. Rabinovich, Nisioi, Ordan and Wintner 2016). Finally, the texts were sliced into chunks of 500 sentences, a number chosen to ensure that their length did not interfere with the analyses (cf. Volansky, Ordan and Wintner 2015). The last slice of each subcorpus was deleted, as these were shorter than 500 sentences and could, therefore, have skewed the results. The ITrs were not divided into training and test subcorpora but only lemmatized because in this study they were studied one by one.

Table 1. The texts in the corpus according to their provenance and language variant.

Language variant	Texts from CTF	Texts from InterCorp	Solicited texts	Scanned texts	Total
De–Fi	2	1	3	0	6
En–Fi	20	16	0	0	36
Fi–Fi	27	25	0	0	52
Fr–Fi	2	1	4	0	7
Gr–Fi	0	0	0	7	7
Sv–Fi	1	3	10	0	14
ITr	0	0	0	13	13
Total	52	46	17	20	135

Table 2. The number of texts and chunks of 500 sentences (lemmatized, shuffled, and sliced) in the training and test subcorpora by language variant.

Subcorpus		No. of texts	No. of chunks of 500 sentences
De–Fi	training	4	38
	test	2	26
En–Fi	training	25	323
	test	11	146
Fi–Fi	training	36	373
	test	16	122
Fr–Fi	training	5	39
	test	2	14
Gr–Fi	training	5	69
	test	2	24
Sv–Fi	training	10	174
	test	4	66

3 Methods

The analyses were done in R using the stylo package (Eder, Rybicki and Kestemont 2016). The main features used in this study included the functions stylo(), with which cluster analysis can be performed, and classify(), which provides supervised methods, such as support vector machines (SVM).

The analyses were based on the frequencies of lemmatized words (most frequent words [MFW]). This means that first, the frequencies of each word (or, in this study, of their lemmatized forms) in the whole (sub)corpus were calculated, and the words were listed from the most to the least frequent. Then, the word frequencies of each individual text were calculated and normalized with z-scores. For example, when a cluster analysis with 100 MFW is run, the first 100 words in the list prepared in the first step are the basis of the analysis: the clustering is based on the frequencies of these 100 words in each individual text. The experiment can also be set to repeat with 30–100 MFW and increases of 10, for example; in this case the test will be done with 1–30 MFW, 1–40 MFW … 1–100 MFW.

An MFW-based analysis is often done by leaving out content words and using only function words in order to fade out topic-specific influences (cf. Grieve 2016; Rabinovich, Nisioi, Ordan and Wintner 2016). There are no widely acknowledged function word lists for Finnish, but a similar effect can be created by using only the words that appear in all the texts; this is done in stylo by setting culling to 100%.

In Finnish, these are words such as *olla*, *ja*, *hän*, *ei*, and *minä* (*to be*, *and*, *s/he*, *no*, and *I*). The MFW function could also be used with other feature sets, such as part-of-speech grams, but because the scanned texts (which include all the Gr–Fi texts and ITrs) have not been cleaned, the accuracy of annotation could distort the results.

Two types of analyses were performed. In the unsupervised cluster analysis, the algorithm clustered the most similar texts together, forming a hierarchical dendrogram that visually illustrated which texts had the most similar MFW profiles. The supervised SVM classifier had two phases. In the training phase, the classifier was given text sets A, B, C … n. It studied their features (in this case, the MFW) and produced a profile for each text set. In the testing phase, the classifier was given text X. It studied its features, produced a profile for it, and compared the profile to those of A, B, C … n to decide which of them was the closest match.

3.1 Distinguishing translations and non-translations

To establish that non-translated Finnish can be distinguished from translated Finnish, a set of three experiments using a SVM classifier was done.

First, experiments were done with 50 chunks of lemmatized, shuffled, and sliced non-translated Finnish and 50 chunks of translated Finnish (consisting of ten chunks of De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi each) in both the training and the test sets. The test run with SVM, with 10–100 MFW at 10-word increases, yielded a general attributive success of 76.4%, meaning that the algorithm correctly identified 76.4% of the chunks as (non)translations. The best result, 97% attributive success, was obtained with 30 MFW; here, the erroneous attributions (three chunks) were non-translated Finnish that were falsely identified as translated Finnish.

Then, to make the experiment as robust as possible, the test was repeated with as many chunks as there were available even if this resulted in the number of chunks ranging from 38 (De–Fi) to 373 (Fi–Fi) in the training set and 14 (Fr–Fi) to 146 (En–Fi) in the test set (see Table 2). The experiment was again based on the SVM, but the number of MFW was narrowed down to 15–50 with increases of 2, as the previous experiment suggested that the best results would be obtained somewhere within that range. The results obtained were slightly stronger, and the best result, 99.2% attributive success, was obtained with 21 MFW. As with the previous experiment setting, the chunks that were wrongly attributed were Fi–Fi chunks misidentified as translated Finnish.

Finally, since the maximum amount of training data seemed to provide the strongest results, the last set of experiments was done, again, using as much training data as possible (Table 2). The test data, however, consisted of the 37 texts in the various test subcorpora (see Table 2) in their full length, which had only been lemmatized (but not shuffled nor sliced) to create a setting resembling real-life conditions, where the method could be applied to full-length texts to verify their (non)translated status. The experiment was done with SVM, with 15–50 MFW in increases of 2. The general attributive success was 81.3%, and the best result, 86.5% attributive success, was obtained with 31 MFW (see Table 3; the one translation erroneously attributed as non-translated Fi–Fi was a Fr–Fi).

Table 3. The confusion matrix of the SVM experiment for distinguishing translations and non-translations with full-length test texts with 31 MFW.

		Attributed		
		Fi–Fi	Tr	Total
Actual	Fi–Fi	12	4	16
	Tr	1	20	21
	Total	13	24	37

3.2 Identifying direct translations' source languages

To take the analysis one step further, a cluster analysis and a SVM analysis to identify the SLs of chunks of translated Finnish were performed. For the cluster analysis, the corpus was tailored to best fit the purpose: all the texts of each language variant (De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi) were put together, lemmatized, shuffled, and sliced into chunks of 500 sentences (once again, the last slices containing less than 500 sentences were deleted) (see Table 4).

Table 4. The number of chunks of 500 sentences (lemmatized, shuffled, and sliced) of translated Finnish by language variant.

Language variant	No. of texts	No. of chunks of 500 sentences
De–Fi	6	64
En–Fi	36	470
Fr–Fi	7	53
Gr–Fi	7	93
Sv–Fi	14	241

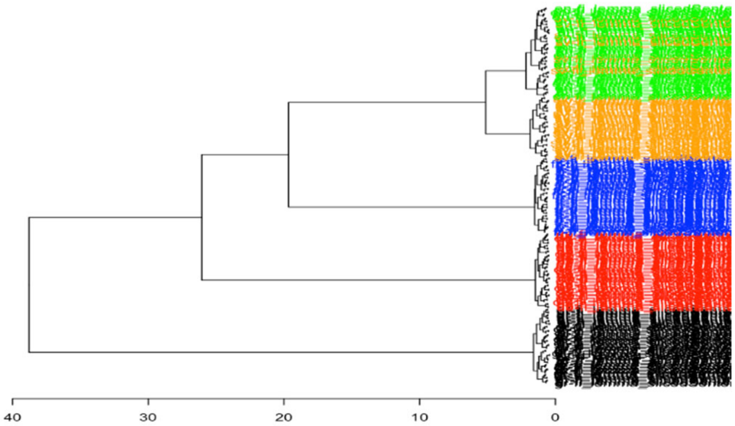

Fig. 1. The dendrogram of the cluster analysis of the translated language variants with 34 MFW; green represents En–Fi, orange Sv–Fi, red De–Fi, blue Fr–Fi, and black Gr–Fi.

In the cluster analysis, 53 chunks (the number of chunks available for the language variant with the smallest number of chunks) of De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi each were used. The clustering was repeated with 2–54 MFW in increments of 2. The dendrogram with 34 MFW (Figure 1) most clearly differentiates the language variants, demonstrating that there is coherence within a group of chunks that were translated from the same SL (they were clustered together in one branch) and variation between the groups of chunks with different SLs (they form different branches), except for Sv–Fi, which paralleled En–Fi to the extent that ten chunks of Sv–Fi were clustered in the En–Fi branch.

After the cluster analysis, a SVM experiment was performed. Here, the pre-manipulated training and test data (Table 2) were used, with the training set consisting of 38 chunks (the number of chunks available for the language variant with the smallest training subcorpus) of each language variant (De–Fi, En–Fi, Fr–Fi, Gr–Fi, Sv–Fi) and the test data of 14 chunks (again, the number of chunks available for the language variant with the smallest test subcorpus) of each language variant. In performing a series of experiments with 2–52 MFW in increments of 2, the general attributive success was 26.3%. The best result, 35.7% attributive success, was gained with 40 MFW. None of the De–Fi nor Fr–Fi chunks were attributed correctly—the Fr–Fi translations were attributed as De–Fi or En–Fi, whereas the De–Fi translations were all attributed as En–Fi; all Gr–Fi translations and roughly one third of the En–Fi and Sv–Fi translations were attributed correctly; no translations were attributed to Fr–Fi (see Table 5).

Table 5. The confusion matrix of the SVM experiment for attributing SLs with 40 MFW.

		Attributed					
		De–Fi	En–Fi	Fr–Fi	Gr–Fi	Sv–Fi	Total
Actual	De–Fi	0	14	0	0	0	14
	En–Fi	0	10	0	0	4	14
	Fr–Fi	7	7	0	0	0	14
	Gr–Fi	0	0	0	11	3	14
	Sv–Fi	0	10	0	0	4	14
	Total	7	41	0	11	11	70

3.3 Experimenting with indirect translations

In this last set of experiments, the aim was to see how the ITrs cluster. The expectation was that they would cluster either with Gr–Fi chunks or with chunks representing translations from their mediating languages. The former would mean that the interference from the ultimate SL carries over through the chain of ITr, and the latter that the interference from the mediating language overrides that from the ultimate SL. For the purpose of this experiment, the 13 ITrs were each lemmatized in their full length and clustered, one by one, with 53 chunks of each language variant (De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi) (Table 4). Since the most discernible SL clusters were previously formed with 34 MFW (Figure 1), this setting was also used to experiment with the ITrs. In other words, the cluster analysis performed in the previous section was repeated 13 times with a different ITr added to each test.

Six of the ITrs clustered with Gr–Fi, two with chunks that represented translations from their mediating language, and the remaining six with neither Gr–Fi nor their mediating language (Table 6). Interestingly, the Fr–Fi, which had previously been misidentified in the SVM-based SL identification, showed a similar tendency here: none of the five ITrs done via French clustered with Fr–Fi chunks (Table 7).

Table 6. The results of the ITr cluster analysis with 34 MFW.

Result	No. of ITrs
Clustered with Gr–Fi	6
Clustered with mediating language	2
Clustered with neither Gr–Fi nor mediating language	5
Total	13

Table 7. The confusion matrix of the ITr cluster analysis with 34 MFW.

		Clustered					
		De–Fi	En–Fi	Fr–Fi	Gr–Fi	Sv–Fi	Total
Assumed	De–Fi	0	0	0	2	1	3
	En–Fi	0	2	0	0	0	2
	Fr–Fi	1	2	0	1	1	5
	Gr–Fi	0	0	0	0	0	0
	Sv–Fi	0	0	0	3	0	3
	Total	1	4	0	6	2	13

4 Conclusions and discussion

The first set of SVM-based experiments proved that translated Finnish can be distinguished from non-translated Finnish: the highest attributive success was 99.2%. An attempt to identify translations' SLs was also made, and although cluster analysis suggested that there is clear coherence within a group of translations from the same SL as well as clear variation between the groups of translations with different SLs, the best attributive success obtained with a SVM classifier was only 35.7%. The unsupervised cluster analysis yielded better accuracies than SVM because it does not make use of separate training and test sets, and, therefore, the variation within each language variant subcorpus is distributed equally to all the chunks.

The poor results with the SVM-based SL identification may be due to insufficient data, as suggested by the fact that none of the chunks of Fr–Fi and De–Fi, the language variants with the least variegated training and test subcorpora, were correctly attributed. SVM works better with more variegated subcorpora: the more variation there is to make the training profiles robust, the higher the attributive success. The need for a varied corpus is a limitation of supervised machine learning. For example, if the training corpus contained translations from one SL only by translator X, the translator's style might override the SL features and become the defining element in the profile created by SVM for the training corpus. If, however, the training corpus also contained translations by translators Y and Z, the translators' individual styles would fade out and the common denominator, the same SL, would become the defining feature of the profile. A similar effect may also explain why the only language variant that was perfectly attributed was Gr–Fi; rather than the SVM identifying features caused by interference from a specific SL, the fact that these texts were not cleaned after the OCR may have left behind a feature that immediately distinguished the Gr–Fi translations from all other language variants. However, cleaning the 20 novels translated from Greek manually was not an option due to time constraints. Similarly, if only one text per author/translator had been allowed in the corpus to increase variation, some of the subcorpora would have become too small for the purposes of this study.

Since the cluster analysis could distinguish the SL variants, the last stage of the study, the SL identification with ITr, was also done using cluster analysis. Six of the thirteen ITrs clustered with Gr–Fi, suggesting that the signal of the ultimate SL carries through the chain of translations to the language of ITrs; this conjecture is bolstered by the fact that only two ITrs (which both had English as their mediating language) clustered correctly with their mediating languages. However, five ITrs did not cluster with either Gr–Fi or their mediating language. Perhaps the language of ITrs is actually mixed, containing traces of both the ultimate SL and the mediating language, and making SL identification impossible when using the language of direct translations as reference data. Should this supposition be correct, it might be possible to distinguish the specific language variants of indirect translations (e.g., Gr–De–Fi, Gr–Fr–Fi). However, sometimes translators consult several source texts in different languages, which could lead to the creation of further language varieties. Alternatively, different passages in the translation could contain different language varieties, in which case a windowing procedure, which focuses on one passage at a time, should be performed to identify the SL passage by passage.

It would be interesting to repeat the SVM-based SL identification experiment with a more robust corpus to see if this would yield better attributive success. If so, then the method should also be tested with ITrs. In any case, since the SVM-based classifier can distinguish between translated and non-translated language, it could be tested with pseudotranslations, that is, texts that pretend to be translations although they are not (Du Pont 2005), to see if the method could be used to expose impostor translations. Ultimately, developing a method to identify translations' SLs could help locate new data for the study of ITr and pseudotranslation. In addition, further studies on these phenomena could provide new information on two specific types of interference: one, in which the interference is fake, as with pseudotranslations, and another where it is mixed, as with compilative ITrs.

References

Assis Rosa, A., Pięta, H. & Bueno Maia, R. (2017). Theoretical, methodological and terminological issues regarding indirect translation: An overview. *Translation Studies 10*(2), 113–132.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 233–250). Amsterdam: John Benjamins.

Baroni, M. & Bernardini S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computin*g *21*(3), 259–274.

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics 13*(3), 411–427.

Corpus of Translated Finnish = Käännössuomen korpus. Käännössuomen sähköinen tutkimusaineisto. Käännössuomi ja kääntämisen universaalit -hankkeessa koostanut Joensuun yliopiston kansainvälisen viestinnän laitos 1997–.

Dollerup, C. (2000). Relay and support translations. In A. Chesterman, N. Gallardo San Salvador & Y. Gambier (Eds.), *Translation in Context* (pp. 7–26). Amsterdam: John Benjamins.

Du Pont, O. (2005). Robert Graves's Claudian novels: A case of pseudotranslation. *Target 17*(2), 327–347.

Eder, M., Rybicki, J. & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *R Journal 8*(1), 107–121.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing 22*(3), 251–270.

Islam, Z. & Hoenen, A. (2013). Source and translation classification using most frequent words. *International Joint Conference on Natural Language Processing*, 1299–1305.

Ivaska, L. (forthcoming). The genesis of a compilative translation and its *de facto* source text.

Ivaska, L. (2018). Three methods to uncover the de facto source language(s) of translations. Poster presented at the European Summer University in Digital Humanities, Leipzig, Germany, 17–27 July.

Ivaska, L. (2016). Uncovering the many source texts of indirect translations: Indirect translations of Modern Greek prose literature into Finnish 1952–2004. Poster presented at the 8th European Society for Translation Studies Congress, Aarhus, Denmark, 15–17 September.

Ivaska, L. & Paloposki, O. (2018). Attitudes towards indirect translation in Finland and translators' strategies: compilative and collaborative translation. *Translation Studies 11*(1), 33–46.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1318–1326.

Lynch, G. & Vogel, C. (2012). Towards the automatic detection of the source language of a literary translation. *Proceedings of COLING 2012: Posters*, 775–784.

Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals: Do they exist?* (pp. 65–82). Amsterdam: John Benjamins.

Rabinonvich, E., Nisioi, S., Ordan, N. & Wintner, S. (2016). On the similarities between native, non-native and translated texts. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016), Berlin, Germany*, 1870-1881.

Straka, M. & Strakova. J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, Vancouver, Canada*, 88–99.

Toury, G. (1995[/2012]). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Ustaszewski, M. (2018). Tracing the effect of pivot languages in indirect translation. In S. Granger, M.-A. Lefer & L. Aguiar de Souza Penha Marion (Eds.), *Using Corpora in Contrastive and Translation Studies Conference, Louvain-la-Neuve, 12–14 September, 2018. CECL Papers 1* (pp. 174–176). Louvain-la-Neuve: Université catholique de Louvain.

Volansky, V., Ordan, N. & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities 30*(1), 98–118.