Quantitative Evaluation of Alternative Translations in a Corpus of Highly Dissimilar Finnish Paraphrases

Li-Hsin Chang, Sampo Pyysalo, Jenna Kanerva, and Filip Ginter
TurkuNLP Group
Department of Computing
Faculty of Technology
University of Turku, Finland
{lhchan, sampyy, jmnybl, figint}@utu.fi

Abstract

In this paper, we present a quantitative evaluation of differences between alternative translations in a large recently released Finnish paraphrase corpus focusing in particular on non-trivial variation in translation. We combine a series of automatic steps detecting systematic variation with manual analysis to reveal regularities and identify categories of translation differences. We find the paraphrase corpus to contain highly non-trivial translation variants difficult to recognize through automatic approaches.

1 Introduction

The study of translation language for Finnish has largely focused on individual linguistic features. The debate on the existence of translation universals sparked the well-developed research line of comparing translated and original language. Examples of such studies include the comparison of nonfinite structures in translated and original Finnish (Puurtinen, 2003; Eskola, 2004), and investigation of subject changes in translations using a French-Finnish parallel corpus (Huotari, 2021). Variation in alternative translations is less studied. Paloposki and Koskinen (2004) qualitatively compare the degree of domestication in language use in Finnish first translations and retranslations. While this study is done qualitatively, several paraphrase corpora with translated language have been released more recently, enabling research from a quantitative prospective. Such corpora include Opusparcus (Creutz, 2018) and TaPaCo (Scherrer, 2020), both constructed automatically using language pivoting and containing Finnish subsets.

Recently, the Turku Paraphrase Corpus has become available (Kanerva et al., 2021), consisting of paraphrase pairs, of which the vast majority are manually selected from the OpenSubtitles[1] dataset. The construction of the paraphrase corpus capitalizes on the fact that many movies and TV shows have multiple independently produced translations. The selection is carried out manually, comparing side-by-side the two lexically maximally distant subtitle versions for each movie or TV show and selecting instances of paraphrases. Upon selection, the candidate pairs are assigned to a category such as *paraphrase in any context* or *paraphrase in this context but not universally*, etc. The Turku paraphrase corpus is substantial in size, with 45,000 manually extracted, naturally occurring paraphrase pairs (a paraphrase pair henceforth refers to two segments of text, each about a sentence long or slightly longer), and a further 7,900 pairs created by editing an extracted pair so as to obtain a fully context-independent paraphrase.

Due to the way in which it was constructed, the corpus is directly applicable to the study of translation language and in particular to the analysis of variation in translation. The unique value of the corpus for this purpose is that it consists mostly of fully manually selected translation variants focused on lexically and structurally dissimilar pairs. These are very difficult to extract automatically: automatic methods can reliably identify only simple variation, while lexically and structurally substantially different pairs are very difficult to automatically distinguish from non-paraphrases, i.e. phrases that are not alternative translations.

In this paper, we will characterize the paraphrase corpus in terms of translation language, focusing especially on the types of variation (e.g. synonym usage, redundancy or verbosity) occurring in the data. Our aim is to establish whether the corpus can be of utility to translation language modelling and machine translation system evaluation. To this end, we will focus on two main ques-

[1] http://www.opensubtitles.org

tions: (a) how easily could the translation pairs be extracted automatically, and (b) what are the main types of variation exhibited by the pairs.

2 Corpus statistics and pre-processing

The full corpus includes 45,000 naturally occurring paraphrases and 7,900 pairs obtained by rewriting a previously extracted example. The source of these paraphrases is in the vast majority of cases alternative translations of subtitles, with a small section originating from news headings. To construct a lexically and structurally diverse paraphrase corpus, the annotators were instructed to only select non-trivial paraphrase candidates, avoiding simple, uninteresting changes such as minor differences in inflection and word order.[2] For the analysis in this paper, we use the training section of the corpus, restricting further exclusively to examples originating from OpenSubtitles. This gives 34,561 naturally occurring paraphrase pairs and 5,445 rewritten paraphrases. Each naturally occurring paraphrase pair in the corpus have a numerical label manually assigned by an annotator from the following set: `4`: universally paraphrase regardless of context, `3`: paraphrase in the given context but not universally, `2`: related but not paraphrase. Additionally, those annotated as `4` can be assigned one or several flags which sub-categorize different types of paraphrases: > or <: universal paraphrase in one direction but not the other, `s`: substantial difference in style, `i`: meaning-affecting difference restricted to a small number of morphosyntactic features. By contrast to the original paraphrases, the rewrites are always full, universally valid paraphrases, i.e. label `4`. The rewriting process strives to change as little of the original sentences as possible: these include simple fixes such as word or phrase deletion, addition or re-placement with a synonym or changing an inflection, while more complicated changes are avoided. The rewrites are thus an efficient way to obtain full paraphrases in terms of corpus creation. The label distribution of the Turku paraphrase corpus subset used for later analysis is shown in Table 1.

For the purpose of the subsequent analysis, we parse the paraphrases using the Turku Neural Parser Pipeline (Kanerva et al., 2018, 2020), a state-of-the-art parser producing POS and morphological tags, word lemmas, as well as dependency trees in the Universal Dependencies scheme (Nivre et al., 2016). We use the model trained on UD_Finnish-TDT v2.7 corpus, which utilizes the pre-trained FinBERT language model in tagging and dependency parsing (Virtanen et al., 2019).[3]

Universal paraphrases	14,986
Label 4	8,578
Label 4s	963
Rewrites	5,445
Context-dependent paraphrases (Label 3 or has <, >, or `i` flags)	24,757
Related but not paraphrase	263
Total	**40,006**

Table 1: Label distribution of paraphrases from the subset of alternative subtitle translations in Turku paraphrase corpus training set.

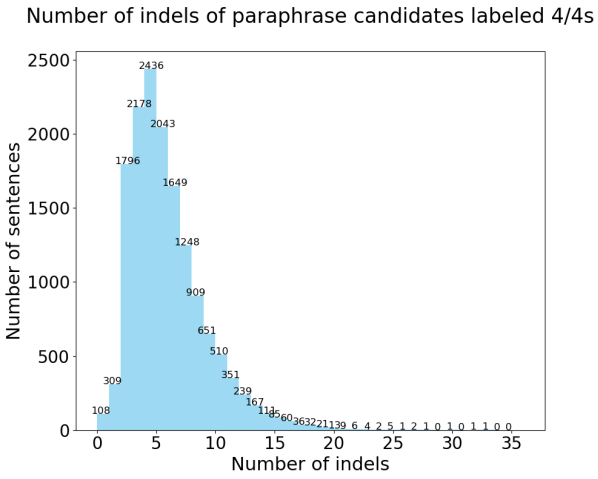

Figure 1: Distribution of the number of lemma indels for universal paraphrases labeled `4`/`4s` including rewrites.

3 Analysis of variation

3.1 Automatic categorization

To investigate and categorize the paraphrase pairs by the form of variation, we calculate the difference in the set of lemmas (i.e. insertions/deletions of lemma, henceforth lemma indels) for each pair, excluding punctuation characters from the analysis. Figure 1 shows the distribution of the number of lemma indels for all universal paraphrases showed in Table 1 (paraphrases with labels `4` and `4s` including rewrites), i.e. all pairs

[2] Finnish has relatively free word order and reordering can be trivially detected automatically.

[3] Model available at `https://turkunlp.org/Turku-neural-parser-pipeline/models.html`

Ratio	Word	Indel	Total
0.45	tosi (really)	64	143
0.41	lakata (stop)	51	125
0.39	ikävä (unfortunate)	55	142
0.38	tahtoa (want)	83	216
0.37	ihan (quite)	145	391
0.35	todella (really)	201	572
0.34	kai (perhaps)	107	311
0.34	aivan (exactly)	117	343
0.34	kyllä (truly)	158	465
0.34	ikinä (never)	127	374

Table 2: Most overrepresented words varying between different translations (minimum occurrence in corpus=50)

equivalent in meaning regardless of their context. As a result of excluding trivial paraphrase candidates, less than 1% (108 pairs) out of 14,986 pairs have zero lemma indels. Such pairs are formed purely by word reordering and/or changes in inflection. We next investigate paraphrase pairs that can be accounted for by automatic synonym substitutions. We combine two resources to build a synonym dictionary for lemmas. The first resource is `Word2Vec` embeddings (Mikolov et al., 2013) for lemmas trained from Suomi24 discussion fora texts[4]. For each lemma, we take at most 15 closest lemmas in the vector space as synonyms using the `gensim` library (Řehůřek and Sojka, 2010). In addition, we supplement our synonym dictionary with Finnish WordNet (Lindén and Niemi, 2014) using the NLTK library (Bird et al., 2009). Out of the 14,878 pairs of paraphrases with lemma indels, 951 pairs (∼6%) have all of their lemma indels accounted by synonyms. An additional 7370 pairs (∼49%) have lemma indels partially accounted by synonyms. The synonym dictionary only takes into account one-to-one synonyms. As a consequence, one-to-many synonyms and phrasal paraphrases are not included.

Table 2 shows the lemmas that are most overrepresented among the inserted or deleted words relative to their overall frequency. We find emphasizers (e.g. *tosi (really)*), particles (e.g. *kyllä (truly)*), auxiliary verbs, other functional words, and a small number of very common synonym pairs among the most frequently varying words.

To further focus on meaningful variation, we disregard all words with a dependency relation deemed functional in the Content-Word Labeled Attachment Score (CLAS) (Nivre and Fang, 2017), which is developed to evaluate dependency parsing with focus on content-bearing words.[5] After disregarding these functional words, we are able to account for the variation in a further 82 paraphrase pairs. All of the above mentioned findings are summarized in Table 3. As the variation in 13,608 pairs (i.e. full 90% of the data) is not accountable by using the above automatic categories, we characterize these manually.

4/4s	14986
Word reordering	1
Same lemma, same order	27
Same lemma, different order	80
CLAS	82
Synonym	945
Synonym + CLAS	243
Others	13608

Table 3: Automatic classification of universal paraphrases labeled `4/4s` including rewrites.

3.2 Manual categorization

In the manual categorization, we sample 100 paraphrase pairs among those paraphrases where the variation is not fully explainable using the automatic metrics defined above. Each paraphrase pair is annotated in terms of 8 different variation categories: *word-to-word*, *word-to-phrase* and *phrase-to-phrase* synonyms indicating a straightforward single word synonym replacement, a single word replaced with a synonymous phrase, or a phrase replaced with a synonymous phrase, *redundancy or verbosity* for including additional words not strictly essential for the meaning, *explicit pronouns* for explicitly including pronouns visible otherwise in the verb inflection, *emphasizer* for including additional emphasis words (such as very), *figurative language/idioms*, and *uncertainty or hedging* where both statements express hedging with different markers.

For each paraphrase pair a set of categories explaining the variation is annotated. In Table 4 we

[4] `dl.turkunlp.org/finnish-embeddings/finnish_s24_skgram_lemmas.bin`

[5] These dependency relations are `aux` (auxiliary), `aux:pass` (passive auxiliary), `case` (pre/postposition), `cc` (coordinating conjunction), `clf` (classifier), `cop` (copula), `det` (determiner), `mark` (marker), `punct` (punctuation), `cc:preconj` (preconjunct), and `cop:own` (copula in possessive clauses).

Category	Count	Ratio
Word-to-word synonym	61	34%
Word-to-phrase synonym	33	18%
Phrase-to-phrase synonym	22	12%
Redundancy or verbosity	21	12%
Explicit pronouns	16	9%
Emphasizers	14	8%
Figurative language/idioms	9	5%
Uncertainty or hedging	3	2%

Table 4: Manual analysis results

plot the frequency of each category, showing the straightforward single word synonym replacement being by far the most frequent category, occurring in 61% of the paraphrase pairs. However, albeit word-to-word replacement being frequent, it rarely accounts for the whole variation in the pair. Only 12% of the paraphrases include word-to-word synonyms as sole variation category, other instances occurring in combination with at least one additional variation category.

3.3 Amount of Non-elementary Variation

We measure the proportion of non-elementary variation in the alternative translations in terms of percentage of text (in terms of alphanumeric characters) in the manually extracted paraphrase pairs, out of the total amount of the source material that the annotators processed. The proportion is 15.8%, meaning that approximately every sixth line was considered to be dissimilar in an interesting manner by the annotators, enough to be included in the paraphrase corpus. The remaining 84% of the text is reported by the corpus creators to be for the most part elementary variation, text without correspondence in the other subtitle version, conflicting erroneous translations, and rarely pairs that are meaningless without deep understanding of their broader context.

3.4 Language pivoting

To establish the proportion of the manually extracted paraphrase pairs that could be identified through their source text, as well as to establish the feasibility of automatically aligning the paraphrase pairs with their English source, we use the OpenSubtitles section of the OPUS machine translation dataset and identify those pairs in our dataset that have at least one common English source segment in the English–Finnish OpenSubtitles section of OPUS. We normalize both Finnish and English texts by lowercasing and dropping all non-alphanumeric characters so as to maximize the recall.

Such language pivoting is a common technique for mining cases of translation variation. Language pivoting targets candidates, where the same source-language segment is translated into two different target-language segments, using a corpus of aligned bilingual document pairs. The candidates are typically further filtered by various means to remove spurious alignments and other pairs which are not equivalent in meaning, despite sharing the same aligned source-language segment.

We find that 2,136 pairs were matched, a mere 6% of all categories of paraphrase in the corpus (barring rewrites). Full 94% of the paraphrase pairs cannot be reached through simple language pivoting at least on the level of full segments. Further, while the average length of texts found through pivoting is 3.8 tokens, the average length of texts in the data is 8.4 tokens. The pivoting thus unsurprisingly biases towards short segments, that are more likely to be appropriately aligned and identified. Clearly, in order to align the paraphrase pairs with their (mostly English) source, a manual annotation step will be necessary.

4 Discussion, Conclusions and Future Work

In this paper, we have presented a quantitative analysis of a large, manually extracted paraphrase dataset from the point of view of translation language, and especially its non-elementary variation. Our findings are two-fold. Firstly, we demonstrated that in the case of OpenSubtitles — a very widely used corpus in machine translation — the proportion of non-elementary variation in alternate translations is relatively small, at 16% of the text. Secondly, we have shown that the paraphrase corpus contains highly non-trivial translation variants that are difficult to account for through simple heuristics and can thus serve for further study in translation language without biasing the results towards simpler examples that can be identified automatically.

The corpus in its current form can serve as a resource for evaluating robustness of different evaluation metrics. Quora Question Pairs (QQP)[6] and

[6] data.quora.com/First-Quora-Dataset-\
Release-Question-Pairs

the QQP subset of Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019) have been used to evaluate the robustness of machine translation and image captioning metrics (Zhang et al., 2020). QQP is a collection of question headings from the Quora forum labeled as either duplicate or not, while PAWS is an adversarial dataset automatically generated from QQP and Wikipedia to contain highly lexically similar paraphrases and non-paraphrases. Based on our findings, the Turku paraphrase corpus serves as an interesting resource to be used in a similar manner to evaluate metric robustness. An obvious direction for future work is to align, through a combination of heuristics and manual annotation, the paraphrase pairs with their English source. This would result in a test set suitable for evaluation of machine translation systems in terms of their rephrasing ability, as well as for research on MT system evaluation methodology in presence of substantial rephrasing.

Acknowledgments

The research presented in this paper was partially supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG). The research was also supported by the Academy of Finland and the DigiCampus project. Computational resources were provided by *CSC — the Finnish IT Center for Science*. We thank Veronika Laippala for her advice from a linguistic point of view.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Mathias Creutz. 2018. Open Subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sari Eskola. 2004. Untypical frequencies in translated language: A corpus-based study on a literary corpus of translated and non-translated Finnish. In Anna Mauranen and Pekka Kujamäki, editors, *Translation Universals: Do they exist?*, pages 83 – 99. John Benjamins Publishing Company.

Léa Huotari. 2021. *Effet du prototype sur le changement de sujet en traduction : Étude d'un corpus bidirectionnel littéraire français↔finnois*. Ph.D. thesis, University of Helsinki.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Finnish paraphrase corpus. In *Proceedings of the 23nd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2020. Universal Lemmatizer: A sequence to sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, pages 1–30.

Krister Lindén and Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? An evaluation. *Language Resources and Evaluation*, 48(2):191–201.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden.

Outi Paloposki and Kaisa Koskinen. 2004. A thousand and one translations: Revisiting retranslation. In Gyde Hansen, Kirsten Malmkjaer, and Daniel Gile, editors, *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*, pages 27 – 38. John Benjamins Publishing Company.

Tiina Puurtinen. 2003. Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals? In Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson, editors, *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, pages 141 – 154. Brill.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 6868–6873.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

A Example instances of manual analysis categories

	Translation$_1$	**Translation**$_2$
Word - word	Vasta ammuttu	Ammuttu hiljattain
	Olen pistämättömän hygieeninen.	Olen moitteettoman hygieeninen.
	Etkö mennyt poliisin luo?	Et mennyt poliisin puheille?
	[...] on luultavasti uusi identiteetti.	[...] on varmasti uusi henkilöllisyys.
Word - phrase	Anteeksi odotus.	Anteeksi, että kesti.
	En edes osaa näytellä.	En edes tiedä miten näytellä.
	On niin paljon valinnanvaraa.	On niin paljon mistä valita.
	Useimmat teistä tietävät [...]	Suurin osa teistä tietää, [...]
Phrase - phrase	Andrew ehti ensin.	Andrew oli vain nopeampi.
	Iän myötä [...]	Mitä vanhemmaksi tulin, sitä [...]
	Miksi hän tekee niin? Etkö ole utelias?	Etkö halua tietää miksi hän tekee niin?
	kuuluuko seuralaisennekin tilin osakkaisiin?	Kuuluuko tili myös seuralaisellenne?
Figurative	Olen täysin hereillä, [...]	Olen pirteä kuin peipponen [...]
	Ole nyt vain hiljaa.	Pidä nyt vain pääsi kiinni.
	Teitkö sen tasataksesi tilit?	Teitkö sen päästäksesi tasoihin?
	Tiedä häntä.	En minä tiedä.
Emph.	Jopa runoja.	Runojakin.
	En tiennyt koko säännöstä.	En edes tiennyt säännöstä.
	Mitä täällä tapahtui?	Mitä ihmettä täällä on tapahtunut?
	[...] näen asiat selvemmin.	[...] näen kaiken aina selvemmin.
Verbosity/ redund.	Voin kertoa teille, että [...]	Se mitä voin kertoa teille, on että [...]
	Se, ketä etsit, on kuollut!	Se ihminen jota etsit on kuollut!
	Mihin voin laittaa tämän? Pedille.	Minne voin laskea tämän? Voit laittaa sen sängylle.
	Hae ensiapupakkaus vessan kaapista.	Hae ensiapupakkaus. Se on vessan kaapissa.
Hedge	[...] herättävätkö ne liikaa huomiota.	[...] että ne saattavat kiinnittää liikaa huomiota.
	Vihaan [...] luultavasti ehkä enemmän [...]	Vihaan [...] ehkä enemmänkin [...]
	Lapset taisivat [...]	Näyttää siltä, että lapset [...]

Table 5: Examples of manual analysis categories. English translations in Table 6.

	Translation$_1$	**Translation**$_2$
Word - word	Recently shot	Just shot
	I am spotless clean	I am perfectly clean
	Didn't you approach the police?	Didn't you talk to the police?
	[...] is likely a new identity.	[...] is surely a new ID.
Word - phrase	Sorry the wait.	Sorry, that it took long.
	I can't even perform.	I don't even know how to perform.
	The choice is so varied.	The choice is very broad.
	Most of you know [...]	The biggest part of you know, [...]
Phrase - phrase	Andrew made it there first.	Andrew was simply faster.
	With age [...]	The older I became, [...]
	Why is he doing so? Aren't you curious?	Don't you want to know why he is doing so?
	Does your colleague also belong among the stock holders?	Does the stock belong also to your colleague?
Figurative	I am fully awake, [...]	I'm astir as a bird [...]
	Be quiet now.	Keep your mouth shut.
	Did you do it to even the score?	Did you do it to get equal?
	God knows.	I don't know..
Emph.	Quite the poem.	A poem.
	I didn't know of the rule as such.	I really didn't know of the rule.
	What happened here?	What on earth happened here?
	[...] you see things more clearly.	[...] you always see everything more clearly.
Verbosity/ redund.	I can tell you that [...]	What I can tell you is that [...]
	The one you are looking for is dead!	The person you are looking for is dead!
	Where can I put this? On the bed.	Where can I lay this down? You can put it on the bed.
	Fetch the first aid kit from the cupboard in the washroom	Fetch the first aid kit. It is in a cupboard in the washroom.
Hedge	[...] do they attract too much attention.	[...] that they may attract too much attention.
	I hate [...] presumably maybe more [...]	I hate [...] maybe even more [...]
	The kids might [...]	It seems that the kids [...]

Table 6: Examples of manual analysis categories, best-effort translation to English.