


ARTICLE

Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for Finnish

Jenna Kanerva*, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón and Otto Tarkka

TurkuNLP, Department of Computing, University of Turku, Turku, Finland

*Corresponding author. Email: jmnybl@utu.fi

(Received 1 July 2022; revised 21 December 2022; accepted 17 January 2023; first published online 16 March 2023)

Abstract

In this paper, we study natural language paraphrasing from both corpus creation and modeling points of view. We focus in particular on the methodology that allows the extraction of challenging examples of paraphrase pairs in their natural textual context, leading to a dataset potentially more suitable for evaluating the models' ability to represent meaning, especially in document context, when compared with those gathered using various sentence-level heuristics. To this end, we introduce the Turku Paraphrase Corpus, the first large-scale, fully manually annotated corpus of paraphrases in Finnish. The corpus contains 104,645 manually labeled paraphrase pairs, of which 98% are verified to be true paraphrases, either universally or within their present context. In order to control the diversity of the paraphrase pairs and avoid certain biases easily introduced in automatic candidate extraction, the paraphrases are manually collected from different paraphrase-rich text sources. This allows us to create a challenging dataset including longer and more lexically diverse paraphrases than can be expected from those collected through heuristics. In addition to quality, this also allows us to keep the original document context for each pair, making it possible to study paraphrasing in context. To our knowledge, this is the first paraphrase corpus which provides the original document context for the annotated pairs.

We also study several paraphrase models trained and evaluated on the new data. Our initial paraphrase classification experiments indicate a challenging nature of the dataset when classifying using the detailed labeling scheme used in the corpus annotation, the accuracy substantially lacking behind human performance. However, when evaluating the models on a large scale paraphrase retrieval task on almost 400M candidate sentences, the results are highly encouraging, 29–53% of the pairs being ranked in the top 10 depending on the paraphrase type. The Turku Paraphrase Corpus is available at github.com/TurkuNLP/Turku-paraphrase-corpus as well as through the popular HuggingFace datasets under the CC-BY-SA license.

Keywords: Paraphrasing; Corpus annotation; Finnish; Paraphrase modeling

1. Introduction

Restating the same meaning in different wording, that is paraphrasing, occurs naturally in human communication, either by the same speaker repeating the message multiple times with different words, or by multiple speakers conveying the same message in different places. While a strict definition of a paraphrase requires the two statements to convey exactly the same meaning, often in natural language processing (NLP) and computational linguistics studies some form of a practical definition is adopted, requiring only having approximately the same meaning. The degree to



which the strict definition is relaxed differs across the various works that address paraphrasing (Bhagat and Hovy 2013).

In NLP, paraphrasing poses interesting challenges in the context of different natural language understanding and generation tasks such as machine translation, machine reading, plagiarism detection, question answering, and textual entailment (Mehdizadeh Seraj, Siahbani, and Sarkar 2015; Altheneyan and Menai 2019; Soni and Roberts 2019). The large, pre-trained language models that have recently become the methodological backbone of NLP have brought about a distinct shift towards more meaning-oriented tasks for model fine-tuning and evaluation. A typical example of such language understanding tasks is entailment detection, with the paraphrase task raising in interest recently, naturally depending on the availability of datasets for the task. Existing paraphrase corpora are typically either large and automatically constructed, or relatively small and manually annotated. Whereas manually annotated corpora are often too small for language model fine-tuning, automatically gathered larger datasets may introduce unwanted bias towards shorter paraphrases with higher lexical similarity due to the corpus-creation heuristics. Moreover, the manually annotated examples are often, although not always, sampled from a larger set of automatically gathered set of examples, carrying over any biases present in the automatic selection heuristics. In view of this situation, there is a need for paraphrase corpora of suitable size for language model fine-tuning, with high quality paraphrases that facilitate language understanding without reliance on surface lexical cues.

In this work, we set out to create a paraphrase corpus for Finnish, specifically aiming at producing a dataset not biased towards simple pairs that can be identified through a simple heuristic. Further, we aim to create a dataset sufficient in size for model training. Our primary motivation is to equip Finnish NLP for research and applications in natural language understanding.

To this end, we develop and apply an extraction protocol for manually collecting text segments that constitute true paraphrases from different paraphrase-rich text sources. Seeing that manual effort is best focused on searching for positive examples of paraphrases, we use automatic extraction of negative paraphrase candidates so as to obtain a dataset suitable for paraphrase classification model training. The concentration of effort on collecting true paraphrases strives for effective usage of the annotation person-months, as nonparaphrases can be more easily collected automatically. In addition, it is a more clearly defined task for the annotators to extract “paraphrases” than to extract “related segments that are not paraphrases”.

Importantly, during the manual paraphrase extraction, the position of the statement in the original source document is stored together with the extracted paraphrase pairs, allowing us to evaluate paraphrases in their natural document context, distinguishing between paraphrases in the given context compared with all possible contexts. To our knowledge, this property sets our work apart from other paraphrase corpora, as it is the first large-scale corpus of sentential paraphrases including manual paraphrase candidate extraction or document context information for the paraphrase pairs.

Together with the dataset, we also examine several paraphrase models trained on the data, as well as include a large-scale paraphrase mining evaluation, where we test how accurately the paraphrase models are able to identify the correct paraphrase pairs when hidden among almost 400M candidate sentences.

The paper is structured as follows. First, we describe the related work in paraphrasing in Section 2. In Sections 3, 4 and 5, we present the overall annotation workflow separated into three phases: heuristic retrieval of related document pairs from different text sources, manual paraphrase candidate extraction from these document pairs, and manual annotation of the extracted candidates. In Section 6, we present the corpus statistics and evaluation, and in Sections 7 and 8, we describe the semi-automatic methods for extracting closely related but negative paraphrase candidates and provide experimental results on both paraphrase classification as well as on paraphrase mining.

2. Related work

Several paraphrase corpora exist, greatly varying in terms of size, extraction methods used, and whether and to what degree the paraphrase pairs undergo manual verification. While most of the paraphrasing studies are carried out on English, paraphrase corpora exist for other languages as well. In addition, a few multilingual paraphrase resources exist. Next, we will review the most relevant work on building paraphrase resources.

2.1. Paraphrase datasets for English

There are numerous English paraphrase datasets in existence. Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett 2005) contains 5.8K paraphrase pairs automatically extracted from an online news collection. Heuristics to identify candidate document pairs and candidate sentences from the documents are used for the extraction, followed by filtering by classifier and finally manual binary annotation using labels (paraphrase or not). Twitter URL Corpus (TUC) (Lan *et al.* 2017) is a collection of 52K paraphrase pairs extracted based on shared URLs in news-related tweets. All pairs are manually labeled to be either paraphrases or nonparaphrases. ParaSCI (Dong, Wan, and Cao 2021) contains 350K automatically extracted paraphrase candidates from ACL and arXiv papers. The extraction heuristics consider term definitions, citation information, and sentence embedding similarity. The paraphrase candidates are automatically filtered without manual labels. ParaNMT-50M (Wieting and Gimpel 2018) contains over 50M sentential paraphrase candidates automatically generated by machine translating the Czech sentences from Czech-English parallel corpora to English. PARADE (He *et al.* 2020) is a collection of 10K paraphrase pairs collected from online user-generated flashcards for computer science related concepts. Definitions for a given term are clustered before in-cluster candidate extraction to reduce candidate selection noise. The candidate examples are subsequently manually assigned labels based on a four-label scheme. Quora Question Pairs (QQP)^a is a collection of question headings from the Quora forum marked with either duplicate or not. Though the QQP dataset is comparatively large (404K pairs) and includes manual labels, the labeling is not originally intended for paraphrasing nor guaranteed to be perfect by the dataset providers. Additionally, Federmann, Elachqar, and Quirk (2019) evaluated different methods for paraphrase dataset generation on 500 English source sentences. These methods include monolingual human paraphrasing as well as translation roundtrip using both human and machine translation on different intermediate languages, but unfortunately the resulting dataset does not seem to be publicly available.

2.2. Other monolingual datasets

Monolingual paraphrase datasets have been constructed for many languages other than English, for instance Chinese, Japanese, Punjabi, Russian, and Turkish. The Phoenix Paraphrasing Dataset,^b released by Baidu, consists of 500K Chinese paraphrase candidates that are short segments of queries. The dataset is created by first collecting seed paraphrase candidates to train a model, which is then used to generate more candidates. The generated pairs are subsequently filtered by a paraphrase recognition model. Shimohata *et al.* (2004) build a Japanese paraphrase corpus containing 683 paraphrase pairs to simplify long spoken-language sentences into machine translation-suitable forms. The paraphrases are travel conversations and their human-paraphrased versions. The paraphrasing strategies are removal of unnecessary redundancy, segmentation of long sentences, and summarization. Arwinder Singh (2020) automatically create a paraphrase dataset for Punjabi with phrasal and sentential paraphrase candidates. They

^a<https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>.

^b<https://ai.baidu.com/broad/introduction?dataset=paraphrasing>.

cluster news headings and articles on the same event from the same day and extract paraphrase candidates with high vector similarity. Nearly 115K phrasal and 75K sentential paraphrase candidates are automatically collected. Manual binary categorization of 1000 pairs from each type shows 88% accuracy for phrasal and 70% for sentential paraphrase candidates. ParaPhraser (Pivovarova *et al.* 2018) is a Russian corpus created through automatic candidate extraction of news headlines from Russian news agencies followed by crowd-sourced manual annotation. It includes over 7K paraphrase pairs classified into nonparaphrases, near-paraphrases, and precise-paraphrases. Due to it not being of sufficient size for text generation, the ParaPhraser Plus dataset (Gudkov, Mitrofanova, and Filippskikh 2020) has been gathered to enable text generation, with over 56M sentential paraphrase candidates. ParaPhraser Plus is created by automatically clustering news headlines by events over a 10-year period and enumerating all pairs of sentences in a cluster. The Turkish Paraphrase Corpus (TuPC) (Eyecioglu and Keller 2018) contains 1002 paraphrase pairs hand-picked from a pool of automatically paired sentences. The automatic pairing involves all-by-all sentence comparison and heuristic filtering based on length and word overlap of sentences from crawled news articles. All selected sentences are manually assigned a numeric label between 0 and 5 quantifying their degree of paraphrase.

2.3. Multilingual paraphrase datasets

Automatic paraphrase recognition oftentimes relies on language pivoting of multilingual parallel datasets. Pivoting is based on the assumption that identical translation possibly entails a paraphrase, and thus use sentence alignments to recognize potential different surface realizations of an identical or near-identical translation. Multilingual paraphrase datasets automatically extracted by language pivoting include Opusparcus (Creutz 2018) and TaPaCo (Scherrer 2020). Opusparcus (Creutz 2018) contains paraphrases for 6 languages and TaPaCo (Scherrer 2020) 73 languages, both including also a Finnish subsection. Opusparcus contains automatically extracted candidate paraphrases from alternative translations of movie and TV show subtitles. While all of the paraphrase candidates are automatically extracted, each language has a manually annotated subset of a few thousand paraphrase pairs. TaPaCo consists of paraphrase candidate pairs automatically extracted from the Tatoeba dataset,^c a multilingual crowd-sourced database of sentences and translations thereof. The paraphrase candidates are assigned into “sets” rather than pairs, and sentences in a set are considered paraphrases of one another. The dataset does not have any manual annotation. Another multilingual paraphrase collection also extracted through language pivoting is Paraphrase Database (PPDB) (Ganitkevitch, Van Durme, and Callison-Burch 2013). Unlike the previously mentioned corpora containing sentential paraphrase candidates, PPDB include only lexical, phrasal, and syntactic paraphrase candidates collected automatically. PPDB has an English collection and a multilingual expansion that includes Finnish (Ganitkevitch and Callison-Burch 2014); however, most of the Finnish candidates in PPDB are just different inflectional variants of the same lexical items.

2.4. Resources for Finnish

The Turku Paraphrase Corpus introduced in this paper, the first large-scale, manually annotated paraphrase corpus for Finnish, includes 91,604 manually extracted and labeled paraphrases with an additional 13,041 human-made rephrasing of statements. While the first incomplete version of the corpus was released in Kanerva *et al.* (2021b), the current work extends the contributions into multiple directions: (1) the corpus size is doubled from the first release, (2) the text sources used to gather the paraphrases are extended from alternative subtitles and news headings to include also news articles, university student essays, translation exercises made by university students, as

^c<https://tatoeba.org/eng/>.

well as messages from an online discussion forum, (3) each manually extracted paraphrase is distributed together with the original document context to allow studies on paraphrasing in context, (4) in addition to manually extracted and labeled paraphrases, an automatically extracted subset of the corpus that contains related nonparaphrase segments is provided to support paraphrase classification.

Apart from the early release of the Turku Paraphrase Corpus, prior to this work only two resources of sentential paraphrases were available for Finnish, the two multilingual datasets Opusparcus and TaPaCo as mentioned above. Opusparcus dataset provides 3700 manually annotated paraphrase pairs for Finnish with an additional release of automatically scored and filtered candidates with different quality threshold ranging from 480K to few million candidates. TaPaCo dataset includes 12K paraphrase candidates for Finnish without any manual verification. A more detailed comparison of these two datasets and our corpus is given in Section 6.2.

3. Text sources for paraphrase extraction

One of the core questions we set out to address in this work is that of bias in paraphrase candidate selection. Here, we specifically want to avoid using heuristics as an initial candidate selection step so as to ensure that the resulting dataset also contains “difficult” examples that would be missed by heuristic selection. To this end, we rely on manual paraphrase extraction, where an annotator receives two related text documents presented alongside each other, and extracts all segments which can be considered as nontrivial paraphrases from the document pair (more details of the actual extraction work is given in Section 4.1). Therefore, in order to obtain sufficiently many paraphrases for the person-months we are able to spend, the text sources used in manual extraction need to be paraphrase-rich, that is have a high probability for naturally occurring paraphrases. Such text sources include for example independently written news articles reporting on the same event, alternative translations of the same source material, different student essays and exam answers to the same assignment, related questions with their replies in discussion fora, and other sources where one can assume different writers using distinct wording to state similar meanings.

We aim to strike a balance between sampling as many text sources as possible, optimizing the usage of person-months available for annotation, and the practical need to reach the goal of 100,000 paraphrase pairs set in the project plan based on which this work was funded. We utilize five different text sources: (1) alternative Finnish subtitles for the same movies or TV episodes, (2) news headings and articles discussing the same event in two different Finnish news sites, (3) different messages with identical title and sub-forum information from a popular Finnish discussion forum, (4) alternative student translations from university translation courses, and (5) student essays answering the same question in university course exams. Next, each text source is described separately introducing the specific methods used to select related document pairs for manual paraphrase candidate extraction.

3.1. Alternative subtitles

OpenSubtitles^d provides a large, vastly multilingual collection of user-contributed subtitles for various movies and TV episodes. The subtitles are available in a large number of languages, and oftentimes there are same-language alternative subtitles for a single movie/episode created independently. These can be viewed as independent translations of the same underlying content and offer an opportunity to make use of the natural variation therein. Through comparing, side-by-side, two alternative subtitle versions of a single movie or TV episode, many naturally occurring paraphrases are likely to be found.

^d<http://www.opensubtitles.org>.

We selected all movies and TV episodes with at least two alternative subtitle versions in Finnish from the database dump of OpenSubtitles2018 obtained through the OPUS corpus (Tiedemann 2012). We measure lexical similarity of alternative subtitle versions by TF-IDF weighted document vectors based on character *n*-grams extracted from within word boundaries. We exclude document pairs with too low or too high document vector cosine similarity values, so as to filter out document pairs with low interesting paraphrase candidate density. This is because a very high similarity often reflects identical subtitles with formatting differences, whereas a very low similarity tends to stem from misalignments caused by incorrect identifiers in the source data and other problems in the data. After this exclusion, the most lexically distant pair is used for paraphrase extraction if there are more than two versions available. For each movie/episode, the two selected subtitle versions are approximately aligned line-by-line using the subtitle timestamps. As we strive to collect paraphrase candidates from as diverse sources as possible, we divide each movie or episode into 15-minute-long segments. For each movie or TV episode, only one or two random segments are used to extract paraphrase candidates. The random selection is intended to prevent accidentally biasing the selection towards typical language used in the beginning of a story.

Altogether, we obtained aligned alternative subtitles for 1700 unique movies and TV series, demonstrating that alternative subtitle versions are surprisingly prolific in OpenSubtitles. We consider movies to be unique items, while episodes from TV series are considered mutually related due to their overlap in plot and characters, resulting in an overlapping in topic and language. After a period of initial annotation, we noticed a topic bias towards certain TV series with large numbers of episodes. We therefore adjusted the number of annotated episodes to be 10 at the highest from each TV series in all subsequent annotation. In total, over 2700 individual movies and TV episodes were used in the corpus construction. Ideally, only one 15-minute segment from each movie or TV episode would be used for candidate extraction, but due to not having enough other paraphrase-rich sources, we conducted a second round of candidate extraction where a second random segment is used after all available movie and TV episodes had been gone through once. The 1300 movies and TV episodes used in the second round were selected based on the number of paraphrase candidate pairs extracted in the first round, the higher the number, the higher the precedence a movie is assigned. In the end, approximately 4100 15-minute-long subtitle segment pairs were used in the corpus construction.

3.2. News articles and headings

We have downloaded news articles through open RSS feeds of different Finnish news sites during 2017–2020, resulting in a substantial collection of news from numerous complementary sources. For the corpus creation, we narrow the data down to two sources: the Finnish Broadcasting Company (YLE) and Helsingin Sanomat (HS, English translation: Helsinki News). The news are aligned using a 7-day sliding window on time of publication, combined with cosine similarity of TF-IDF-weighted document vectors induced on the article body, obtaining article pairs likely reporting on the same event. The parameters of the TF-IDF vectors induction are the same as in Section 3.1. After aligning the candidate documents, article headings and the rest of the article text, referred as article body from now on, are processed separately due to different sampling strategies applied to these. We use a simple grid search and human judgment to establish the most promising region of similarity values in order to avoid candidate pairs with almost identical texts or candidates with similar topic but reporting on different events. While in news article bodies, we strive for balance between too low and too high similarity. In news headings, we target to select maximally dissimilar headings of news articles having maximally similar body texts as the most promising candidates for nontrivial paraphrase pairs. Furthermore, while the promising pairs of article body texts are selected for manual paraphrase extraction, news headings typically include only single sentence-like statements and are thus directly transferred into the paraphrase classification tool skipping the manual extraction phase. A total of approximately 2700 news heading pairs and 1500 article body pairs were used in the corpus construction.

3.3. Discussion forum messages

We hypothesize that different discussion forum messages related to same topics may include a sufficiently large number of naturally occurring paraphrases to justify a manual extraction effort. For example, different thread-starting messages under the same subforum often seek information on the same topic or share related experiences, or different replies to the same message often convey similar reactions. We set out to experiment with thread-opening messages with identical titles posted into the same subforum. We find that while most of the candidate document pairs selected this way are related messages from different authors often discussing similar personal experiences or seeking advice for similar matter. We also noticed a significant number of messages clearly written twice by the same user, with similar overall structure but using a different wording.

We use the public release of the Suomi24 discussion forum^e including over 80M messages posted online between years 2001 and 2017. From the data release, we identify all thread opening messages and align candidate document pairs with identical title and subforum information combined with TF-IDF similarity of messages. Candidate alignments with too low or too high similarity, as well as candidates where the shorter message is merely a subset of the longer one, are filtered out based on preliminary human judgment gridding different similarity threshold values. This produced about 13K candidate message pairs. However, before the actual paraphrase extraction phase, 44% of these were yet discarded in an additional manual annotation step, where candidate document pairs were either accepted or rejected based on the potential estimated by inspecting the first few sentences from both documents. Here, the annotator only quickly verified a reasonable correspondence existing between the document pair without carefully reading the message content. This additional manual annotation step was carried out as we were not able to find an automatic method reliable enough to identify false positives among the candidates. Furthermore, filtering low-quality pairs before the actual paraphrase extraction step was found more efficient than executing filtering and paraphrase extraction simultaneously. In the end, a total of about 7100 accepted message pairs were used in the corpus construction.

3.4. Student translations

Seeing the potential of alternative translations originating from movie and TV episode subtitles, we initiated an attempt to find alternative source material where the same foreign text is translated into Finnish by multiple translators. One potential source of a constant stream of alternative translations is exercised from different language studies and courses, where several students translate the same exercise text. In order to avoid oversimplified short sentences, which one would see in many beginner level courses, we targeted exercises taken from university courses in translation studies where all students have sufficiently good skills and the exercises include translating authentic documents from different sources into Finnish. Such sources would typically include samples of magazine articles, business contracts, advertisements, etc.

We were able to collect 16 unique exercise texts with at least two different student translations. If more than two translations existed for the same source text, at most three different pairs were used in annotation so as to avoid over-extracting repetitive paraphrases, and a total of 28 document pairs were used in the corpus construction. However, the main limitation of student translations is their availability due to data usage regulations.^f

^e<http://urn.fi/urn:nbn:fi:lb-2019021101>.

^fObtaining adequate permissions to use any student produced data involved manual permission inquiries and we found it difficult to motivate the students to give their consent. A long-term collaboration with a translation study program would likely improve this situation.

3.5. University exams

The final text source experimented with is student essays collected from university course exams, where the hypothesis is that all essays answering the same exam assignment will include similar arguments, and therefore, have a high probability for naturally occurring paraphrases. However, the student essays possess the same availability limitations as student translations where the usage of student materials is restricted and requires an explicit written consent.

We were able to collect a total of 34 student exams from three university courses (*Introduction to Language Technology, Corpus Linguistics and Language Technology, and Philosophy of Science and Research Process*). The exams included 24 unique questions or essay assignments for which at least one candidate pair (two alternative essay answers) was available. However, the answers for one assignment often divided into several subtopics because the students were able to select one aspect covered during the course to answer the assignment. The number of unique topics was consequently larger. We therefore processed each unique question/essay assignment/subtopic separately, rather than exams in full. The length of a typical answer varied between few sentences and one full page depending on the assignment. In the end, a total of 190 student answer pairs were used in the corpus construction.

4. Paraphrase candidate extraction

After the heuristic document alignment, the actual paraphrase candidate extraction is based on fully manual work. Next, we describe the paraphrase candidate extraction workflow, evaluate the adequacy of different text sources using several extraction measures, as well as show the distribution of paraphrases originating from different text sources in the final corpus.

4.1. Extraction workflow

Given a document pair extracted from one of the text sources, the manual annotation work begins with manual candidate extraction. In a dedicated candidate extraction tool, an annotator sees both documents simultaneously side-by-side and is instructed to extract all interesting paraphrases from the texts. In order to collect a varying set of nontrivial paraphrases, candidates with simple, uninteresting changes such as minor differences in inflection and word order are avoided during extraction. A paraphrase can be any text segment from few words to several sentences long, and the paraphrase extraction is not restricted to follow sentence boundaries. The two statements in one candidate pair can also be of different lengths, mapping for example one sentence on one side to several on the other side. The annotators are encouraged to select as long continuous statements as possible (rather than splitting them into several shorter ones), nevertheless at the same time avoiding over-extending one of the statements by including a long continuation which does not have a correspondence in its paraphrased version. The annotators are not actively trained to harmonize their personal candidate extraction strategies, since the aim is to include more diverse paraphrase candidates in the corpus, thus minor differences in extraction phase behavior are not considered harmful. The most typical property defining “personal style” in candidate selection was where to place the boundary between interesting and trivial pairs.

When completing the document pair, the annotator marks it finished and continues to the next document pair. After accumulating a reasonable amount of material in the extraction tool, all extracted paraphrase candidates are transferred into a separate paraphrase classification tool, where the annotation work continues as a separate session. Even if these two annotation phases were executed one after the other, the annotators were able to alternate freely between the two tasks in order to keep the working days more varied. Typically, the annotator who extracted the paraphrase candidates also did the labeling in the next phase. However, this is not strictly required and sometimes data is transferred between different annotators due to time constraints.

Table 1. Manual paraphrase extraction statistics for different text sources, where *Documents* refers to the number of document pairs producing paraphrases, *Empty* refers to the percentage of candidate document pairs not producing any paraphrase candidates (all other metrics are calculated after discarding the empty pairs), *Yield* refers to the average number of paraphrase pairs extracted from one document pair, *Coverage* is the total proportion of text (in terms of alphanumeric characters) selected in paraphrase extraction from the original source documents, and *Length* is the average length of the original document in terms of alphanumeric characters. Note that the alternative subtitle statistics are based on the first round of annotations only, where the movie/episode selection is not biased towards high-yield documents, and here one subtitling document refers to a 15-minute segment of a movie/episode

Text source	Documents	Empty (%)	Yield	Coverage (%)	Length
Alternative subtitles	2781	9.2	17.6	17.5	4300
News article bodies	1463	11.4	3.7	24.6	1600
Discussion forum messages	7106	36.7	1.7	22.8	500
Student translations	28	0.0	22.7	75.1	3700
University exams	190	31.6	2.4	25.8	1100

4.2. Extraction statistics

Next, we analyze the different text sources used in the paraphrase extraction in several aspects. When evaluating the adequacy of the text source for the extraction purposes, we find it most interesting to measure how “productive” on average one document pair is. This is measured mainly using two metrics, the percentage of empty documents pairs, where empty refers to a document pair not producing any paraphrase candidates and can therefore be considered “useless” for the corpus construction purposes, as well as paraphrase yield, where yield refers to the average number of paraphrase candidates extracted from a nonempty document pair, where the assumption naturally is that the more one can extract from one document pair, the more time-efficient the extraction process is.

The overall extraction statistics are given in Table 1 separately for all five text sources. In terms of empty document pairs, the percentage varies between 0% and 37%, the two translation-based sources, student translations, and alternative subtitles, include the least amount of empty document pairs. An annotator not being able to extract any paraphrases from the document pair is typically caused by the two documents being lexically too similar and therefore not including interesting paraphrases, or them being topically related without any corresponding parts. In terms of the average yield of paraphrases per pair of documents, the story remains largely unchanged, with the two translation-based sources clearly having the best yield. From student translations, the annotators are able to extract on average 22.7 paraphrase candidates per nonempty document pair and from alternative subtitles the average yield is 17.6 candidates. In the end, it is not surprising that alternative translations yield the most amount of paraphrases as the translation process requires keeping the same basic information as present in the original, while for example in news articles the journalists can more freely select which aspects to report or not to report. Additionally, we were somewhat surprised how many verbatim quotations there were in news articles, where both news agencies clearly used the same reference text and possibly added a paragraph or two of their own text. The average length of the documents also naturally affects the yield, and the source with the worst average yield (discussion forum messages with only 1.7 paraphrase candidates per document pair) also has on average the shortest documents, with many of the discussion forum messages including only 1–2 sentences. In terms of coverage (proportion of the original text selected in paraphrase extraction), the differences are substantially smaller.

The final selection of source materials used for building the Turku Paraphrase Corpus is for the most part determined by two factors: availability and average paraphrase yield in the manual candidate extraction phase. Although the student produced materials were found promising in

Table 2. The number of paraphrase pairs in the released corpus originating from different text sources (rewrites, introduced in Section 5.3, are included in the statistics)

Text source	Paraphrase pairs	% of the corpus
Alternative subtitles	86,170	82%
News	9198	9%
<i>Body text</i>	5450	5%
<i>Headings</i>	3748	4%
Discussion forum messages	8175	8%
Student translations	760	1%
University exams	342	<1%

our experiments, especially the translation exercises which gave the best evaluation numbers in all metrics, the work required to settle legal restrictions on student produced materials prevented any larger-scale utilization of these sources under the scheduling constraints of the project. More groundwork would be required at the university and even national level to ease the usage of such data sources also retrospectively. Additionally, our goal of openly licensing (CC-BY-SA) the produced corpus creates increased complexity compared with a mere academic use in terms of student materials.

The limited amount of student materials left us with three primary text sources, of which alternative subtitles have a clearly better average yield per document pair compared with news articles and discussion forum messages. While news articles and discussion forum messages have better coverage (proportionally more of the source text is extracted), likely due to documents in general being shorter, one could assume the annotator being able to extract the same amount of material by just going through more document pairs. However, the amount of time the annotators spend on one document pair is considerably longer for news articles and discussion forum messages than for alternative subtitles. The main reason for this is that the two alternative subtitling documents are well aligned, while arguments in news articles and discussion forum messages often come in different order, requiring the annotators to scroll up and down in the paraphrase extraction interface in order to find the corresponding arguments. Also, after finding a corresponding argument in both documents, the annotator must yet verify the meaning of the extracted statement in the given context, as one cannot reliably assume the whole document following strictly the same story as in the case of the alternative translations where the source story is guaranteed to be identical. This extraction complexity effect is clearly visible in the weekly paraphrase extraction speed unofficially monitored throughout the project, where the extraction speed halved when switching from alternative subtitles to news articles and discussion forum messages. The extraction speed is thus the second limiting factor when selecting source material for annotation, and consequently, some of the text sources are highly overrepresented in the corpus. The number of paraphrase pairs obtained from different text sources are summarized in Table 2, the alternative subtitles dominating the final dataset with 82%, news texts and discussion forum messages both having a bit less than 10% portion, while both student materials represent only a tiny fraction of the corpus data.

5. Paraphrase annotation

After the candidate extraction, all candidate paraphrases are manually annotated according to the given annotation scheme. Next, we introduce this annotation scheme as well as some of the

more generally interesting annotation guidelines. In the end of the section, we present the overall annotation workflow where the annotators also have an option to provide an additional rewrite of the original paraphrase pair in order to correct small issues in the original candidates.

5.1. Annotation scheme

Many different paraphrase annotation schemes are presented in earlier studies, most commonly falling either into a simple yes/no (*equivalent* or *not equivalent*) as in MRPC (Dolan and Brockett 2005), or a numerical labeling capturing the strength/quality of paraphrases, such as the 1–4 scale (*bad, mostly bad, mostly good* and *good*) used in Opusparcus (Creutz 2018).

Instead of these simple annotation schemes, we set out to capture the level of paraphrasability in a more detailed fashion with an annotation scheme adapted to this purpose. Our annotation scheme uses the base scale 1–4 similar to many other paraphrase corpora, where labels 1 and 2 are used for negative candidates (unrelated/related but not a paraphrase), while labels 3 and above are paraphrases at least in the given context if not everywhere. In addition to base labels 1–4, the scheme is enriched with additional subcategories (flags) for distinguishing a small number of common special cases of paraphrases, which in many respects lie between the labels 4 (universal paraphrase) and 3 (paraphrase in the given context).

5.1.1. Label 4: Universal paraphrases

Label 4 is assigned to cases of a universal (perfect) paraphrase that holds between the two statements in all reasonably imaginable contexts, meaning one can always be replaced with the other without changing the meaning. This ability to substitute one for the other in any context is the primary test for label 4 used in the annotation. Examples of universal paraphrases include:

Tulen puolessa tunnissa.
'I'll be there in half an hour.'
Saavun 30 minuutin kuluessa.
'I will arrive in 30 minutes.' → 4

Voin heittää sinut kotiin.
'I can give you a lift home.'
Pääset minun kydyssäni kotiin.
'You can ride home with me.' → 4

Tyrmistyttävän lapsellista!
'Shockingly childish!'
Pöyristyttävän kypsymätöntä!
'Astoundingly immature!' → 4

With the base scale alone, a great number of candidate paraphrases would fail the substitution test for label 4 and be classified as label 3. This is especially true for any longer text segments which are less likely to express very strictly the same meaning even though conveying the same principal idea. So as to preserve some of the most important such general cases and to avoid overusing the label 3 category with a very diverse set of paraphrases, we introduce flags for finer subcategorization and therefore support a broader range of downstream applications of the corpus as well, since many applications may have different requirements for paraphrases. For instance, if considering rephrasing systems (paraphrase generation), the requirements for paraphrasing are quite strict in order to avoid for example the model learning to introduce additional facts or changing the style into offensive language on its own. On the other hand, in information retrieval,

the paraphrasing is usually more loosely defined, and finding occurrences with more variation is often appreciated. These annotated flags can only be attached to label 4 (subcategories of universal paraphrases), meaning the paraphrases are not fully interchangeable due to the specified reason, but, crucially, are context-independent that is their annotated relationship holds regardless of the textual context, which is unlike label 3. The possible flags are:

Subsumption (> or <). The subsumption flag is for cases where one of the statements is more detailed and the other more general (e.g. one mentioning *a woman* while the other *a person*), with the arrow pointing towards the more general statement. The relation of the pair is therefore directional, where the more detailed statement can be replaced with the more general one in all contexts, but not the other way around. The two common cases are one statement including additional minor details the other omits, and one statement being ambiguous while the other not. If there is a justification for crossing directionality (one statement being more detailed in one aspect while the other in another aspect), the pair falls into label 3 as the directional replacement test does not hold anymore. Examples of paraphrases with directional subsumption are shown below, where the first and second examples are cases of one of the statements including information the other omits (agent in the first example and purpose of the action in the second), while in the third example the latter statement is ambiguous, including both figurative and literal meaning:

Tulit juuri sopivasti.

'You arrived aptly.'

Loistava ajoitus.

'Fantastic timing.' → 4>

Tein lujasti töitä niiden rahojen eteen.

'I worked hard for that money.'

Paiskin kovasti töitä.

'I toiled away.' → 4>

En pysty tähän.

'I cannot do this.'

Tämä on liian suuri pala minulle.

'I'm in way over my head with this one.' → 4>

Style (s). The style flag is for marking tone or register difference in cases where the meaning of the two statements is the same, but the statements differ in tone or register such that in certain situations, they would not be interchangeable. For example, if one statement uses pejorative language or profanities, while the other is neutral, or one is clearly colloquial language while the other is formal. The style flag also includes differences in the level of politeness, uncertainty, and strength of the statements. Examples of paraphrases with different style (examples 1 and 2) and strength (example 3) include:

Helou gimmat!

'Hey, you gals!'

Päivää tytöt!

'Good day, girls!' → 4s

Mistä hitosta tietäisin?

'How the hell should I know?'

Minä en tiedä.

'I do not know.' → 4s

Täällä on aika kylmä ilmapiiri.

'The atmosphere is quite cold here.'

Täällä on jäätävä tunnelma.

'What a chilly mood there is round here.' → 4s

Minor deviation (i). The minor deviation flag marks in most cases minimal differences in meaning (typically *this* vs. *that*) as well as easily traceable differences in grammatical number, person, tense or such in cases where they are determined to have a difference in meaning. Some applications might consider these as label 4 for all practical purposes (e.g. information retrieval), while others should regard these as label 2 (e.g. automatic rephrasing). In cases where the minor change in for example mood or tense does not make a difference in meaning, the minor deviation flag is not marked. However, note that even when these minor differences are accepted, they cannot violate the paraphrasability in the context, for instance replacing the pronoun *minä* 'I' with *sinä* 'you' will not (generally speaking) make a paraphrase, while replacing *minä* 'I' with *me* 'we' can work in some contexts, however, quite rarely. Typical examples of paraphrases with minor deviation flag include:

Tämä laite on epäkunnossa.

'This piece of equipment is malfunctioning.'

Tuo kone on rikki.

'That machine is broken.' → 4i

Teitpä onnisti!

'You (plural) are in luck!'

Oletpa onnekas!

'Aren't you (singular) lucky!' → 4i

Vaimon mukaan hän vihaa tupakointia.

'According to his wife, he hates smoking.'

Hänen vaimonsa sanoo, että hän vihasi tupakan polttamista.

'His wife said that he hated smoking.' → 4i

The flags are independent of each other and can be combined in the annotation (naturally with the exception of > and < which are mutually exclusive).

5.1.2. Label 3: Context dependent paraphrases

Label 3 is a context dependent paraphrase, where the meaning of the two statements is the same in the present context, but not necessarily in other contexts. The common cases include statements, where both are ambiguous in different ways or both include different additional details not strictly necessary for conveying the main message (conflict in the subsumption flag directionality). Examples of context dependent paraphrases are shown below, where in the first example both include different additional details (first statement mentioning *night* while the second including a reference to *you*), while the second and third examples are cases where both statements are ambiguous in different ways or include a use case not covered by the other (e.g. in the third example the *911* can refer to the emergency number or simply be used when counting items, while the *emergency number* is *911* in some countries but not in all):

Miten eiliselä meni?

'How was last night?'

Miten teillä meni eilen?

'How did it go for you yesterday?' → 3

Aion tehdä kokeen.

'I am going to make an experiment.'

Aion testata sitä.

'I am going to test it.' → 3

911.

'911.'

Hätänumero.

'Emergency number.' → 3

5.1.3. Label 2: Related but not a paraphrase

Label 2 means related but not a paraphrase, where there is a clear relation between the two statements, yet they cannot be considered paraphrases in the sense outlined above for labels 4 and 3. Common cases include statements with a significant difference in the main message even if describing the same event, statements with contradictory information present, statements which could be paraphrases in some other context but not in their present context (such examples were very rare), or literal translations of metaphors which fail to communicate the metaphoric meaning in the source text (clumsy but understandable translations do receive label 3). Examples of related statements, which are not paraphrases are shown below, where the first example is topically heavily related and describing the same event but having a different main message, the second example describes the same event but from different point of time (therefore including contradictory information), and the third example includes a literal translation of a metaphor which doesn't make sense after the translation:

Tappion kokenut Väyrynen katosi Helsingin yöhön.

'After suffering defeat, Väyrynen disappeared into the night of Helsinki.'

Väyrynen putoamassa eduskunnasta.

'Väyrynen is in danger of dropping out of the Finnish Parliament.' → 2

Aurassa perjantaina kadonnut 12-vuotias poika löytynyt.

'The 12-year-old boy who went missing in Aura on Friday has been found.'

Poliisi etsii 12-vuotiasta poikaa Aurassa.

'The police are searching for a 12-year-old boy in Aura.' → 2

Olet löytänyt onnen.

'You have found happiness.'

Nyt sinulla on avaimet linnaan.

'Now you have the keys to the castle.' → 2

5.1.4. Label 1: Unrelated

Label 1 is for unrelated candidates, where there is no reasonable relation between the two statements, most likely occurring due to a false positive in candidate selection. If the candidate pair shares only a single proper name while the topic otherwise is different, the candidate is considered unrelated.

Oletteko Sherlock Holmes?

'Are you Sherlock Holmes?'

Riippuu.

'It depends.' → 1

Sipoonranta on Sipoossa, ei Helsingissä.

'Sipoonranta is located in Sipoo, not in Helsinki.'

Sipoonranta hakee taas lisää aikaa rakentamiseen.

'Sipoonranta is again applying for more time for building.' → 1

5.1.5. Label *x*: Skip

If labeling a candidate pair is not possible for another reason, or giving a label would not serve the desired purpose (e.g. wrong language or identical statements), the example can be skipped with the label *x*.

5.2. Annotation guidelines

While each decision in paraphrase annotation must be done based on considering each individual example separately, several systematic differences among the annotators were identified during the annotation process, and comprehensive annotation guidelines were produced to guide the annotation process towards harmonized decisions between different annotators. A total of 17-page annotation manual was produced in collaboration among the annotators, and the guidelines were revised and extended regularly to account for new problematic cases. The full manual is published as a technical report (Kanerva *et al.* 2021a), and some of the most interesting/relevant policies are discussed below.

5.2.1. Syntactic structure

Merely syntactic differences are not accounted in the labeling if they do not change the sentence meaning, even if the difference would make sentence substitution clumsy in some contexts. For example, the lack or inclusion of discourse connectives can make the sentence feel clumsy or isolated from the context, however they barely carry much additional information. The same policy is adapted to for example differing verb tense and mood if the difference does not carry change in meaning. However, if a shift in meaning is noticed it is annotated accordingly.

5.2.2. World knowledge

In certain cases, one of the statements includes additional information which can be seen as world knowledge (facts generally known or knowable by everyone). For example, in the paraphrase pair

Omena on hedelmä, josta valmistetaan mm. hilloa ja mehua.

'An apple is a fruit from which you make jam and juice, among other things.'

Omenasta valmistetaan muun muassa hilloa ja mehua.

'Among other things, jam and juice are made from apples.'

the second statement does not explicitly mention apple being a fruit. However, considering that this is a generally acknowledged fact, which does not contribute to the core meaning, explicitly mentioned additional world knowledge facts are not considered additional information in paraphrase annotation, and therefore, the above-mentioned example would receive label 4 in annotation.

The same principle is adapted for well-known noun modifiers (e.g. permanent titles and descriptive nouns such as *Queen Elizabeth II*, *ski jumping legend Matti Nykänen* or *tech company Microsoft*). However, if the noun modifier is considered to be meant for temporary use only, as many times for example in politics (e.g. *prime minister Sanna Marin*), noun modifiers are considered additional information as it binds the statement into a specific time.

In few cases, the world knowledge principle allows proper name replacement with a common noun phrase, if the entity can be unambiguously individualized from the common noun description. For example, in the paraphrase pair

Ensimmäinen avaruuteen lähetetty suomalaissatelliitti tuhoutui.

'The first Finnish satellite that was launched to space was destroyed.'

Aalto-2 tuhoutui.

'Aalto-2 was destroyed.'

while *The Finnish satellite* could refer to any Finnish satellite, there can be only one “first one”, which then individualizes the noun phrase and the example is annotated with label 4.

5.2.3. Time references

Time references can be either exact (*24.12.1999, in 2020, 16:00 o'clock*) or relative with respect to the current time (*today, last year, in three hours*). When comparing two exact time expressions, the label is 4 if the same amount of information (e.g. day, month, year) is present, but often 4 with a subsumption flag if one of the two is more descriptive and the additional information cannot be considered world knowledge. When comparing two different relative time references with each other (e.g. *in the beginning of the week* and *three days ago*), the label is usually 3 if the time is not further specified elsewhere in the statements. When comparing exact time with relative time, the labels depends on whether the exact time can be considered world knowledge or not. For example, in statements

Matti Nykänen kuoli viime vuoden helmikuussa 55-vuotiaana.

'Matti Nykänen died in February of last year at the age of 55.'

Matti Nykänen kuoli helmikuussa 2019. Hän oli kuollessaan 55-vuotias.

'Matti Nykänen died in February of 2019. He was 55 years old at the time of his death.'

the date of death of a famous person can be considered world knowledge, and the paraphrases can be labeled with label 4> the latter being more general as it can be used in any point of time, while *February of last year* can only refer to the year 2019 in this context and therefore be used only in 2020. When comparing exact time with relative time in the context of events not considered world knowledge, for example in

Rikos tapahtui viime vuoden helmikuussa.

'The crime happened in February of last year.'

Rikos sattui helmikuussa 2019.

'The crime took place in February of 2019.'

the event in question (crime) is not individualized and the exact time cannot be considered world knowledge, therefore the label is 3.

5.3. Annotation workflow

After accumulating a reasonable amount of material in the candidate extraction phase (typically every two to three days), the extracted paraphrase candidates are transferred into a dedicated paraphrase classification tool, where the annotator is able to see all paraphrases extracted from the document pair one by one. In the paraphrase classification tool, the annotator assigns a label for each paraphrase candidate using the above-mentioned annotation scheme. Even if the extracted paraphrases are shown one-by-one in the tool, the full document context is available. In addition

to labeling, the tool provides an option for rewriting the paraphrase pair to be fully interchangeable, universal paraphrases. The annotators are instructed to rewrite paraphrase pairs that are not already label 4, in cases where a simple edit, for example word or phrase deletion, addition, or re-placement with a synonym or changing an inflection, can be easily constructed. Rewrites must be such that the annotated label for the rewritten example is always label 4. In cases where the rewrite would require more complicated changes or would take too much time, the annotators are instructed to move on to the next candidate pair rather than spend time on considering the possible rewrite options.

The classification tool also provides an option to tag examples where the annotator feels unsure about the correct label, the example is particularly difficult, or otherwise more broadly interesting. These examples were discussed in the whole annotation team during daily annotation meetings. The annotators also communicated online, for instance seeking a quick validation for a particular decision.

5.4. Ensuring annotation consistency in early annotations

As the annotation guidelines were revised and extended throughout the corpus annotation, there is the potential of small discrepancies between examples annotated at the very early stage of the project compared with those annotated at the very end. In order to assure the consistency between the revised guidelines and early stage annotations, during the final weeks of annotation several quality assurance rounds were carried out, especially targeting labels whose guidelines changed during early annotation work.

All annotated examples were first divided by labels, and then sorted based on annotation timestamp from earliest to latest. Concentrating on the most problematic labels *s* (flag for style) and *i* (flag for minor deviation), examples including these flags were manually checked and corrected if necessary, starting from the earliest annotations and continuing until the latest guidelines and the annotated examples were in sync, and no systematic errors were noticed anymore. A total of 5.7% of all annotated examples were inspected, of which about 30% were corrected according to the latest guidelines. Time-wise most corrections were dated to the first 2 months of the annotation work.

6. Corpus statistics and evaluation

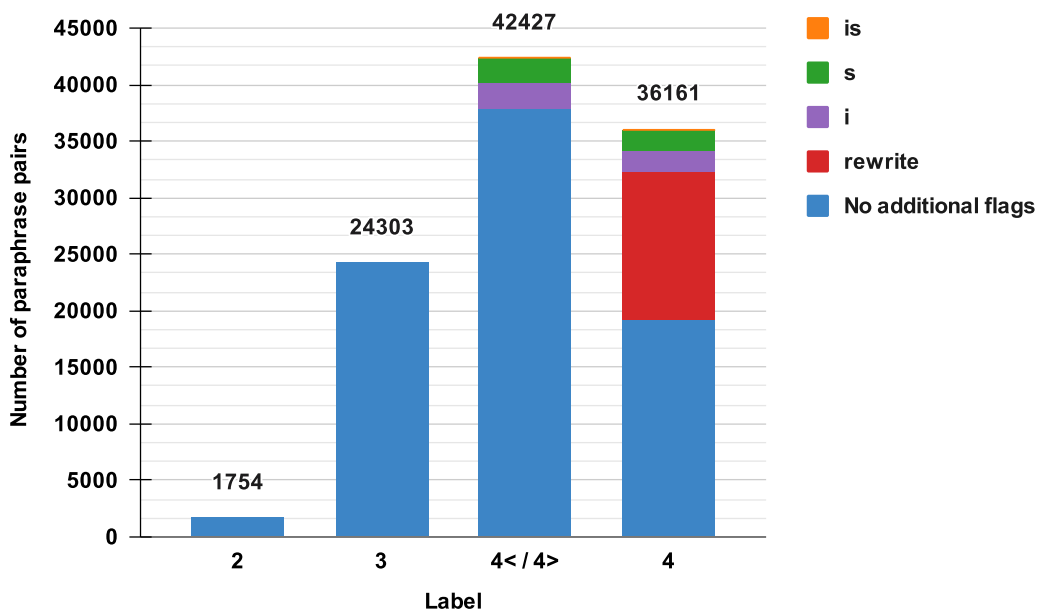
The released corpus is comprised of 91,604 naturally occurring paraphrase pairs extracted from the source documents with an additional 13,041 rewrites, thus resulting in a total of 104,645 manually classified Finnish paraphrase pairs. The data are randomly divided into training, development, and test sections using a 80/10/10 split; however, with the restriction that all paraphrases from the same movie, TV episode, news article, student translation text, or exam question are assigned to the same section. Basic statistics are summarized in Table 3, and the label distribution is shown in Figure 1. As the manual candidate extraction targeted “true” paraphrases, 98% of all annotated paraphrases are classified to be at least paraphrases in their given context (label 3) if not in all contexts (label 4). The number of candidates labeled with labels 1 or *x* is negligible, therefore these are discarded from the corpus release altogether.

6.1. Annotation quality

The annotation work was carried out by six main annotators together with a broader project team supporting their effort. The six annotators used a total of 30 person-months for the corpus construction, where the work includes paraphrase extraction, label annotation as well as other related tasks such as guideline documentation. Each annotator had a strong background

Table 3. The sections of the corpus and their sizes in terms of number of paraphrase pairs

Section	Examples	Rewrites	Total
Train	73,165	10,480	83,645
Devel	9231	1298	10,529
Test	9208	1263	10,471
Total	91,604	13,041	104,645

**Figure 1.** Label distribution in the whole corpus.

in language studies with an academic degree or ongoing studies in a field related to languages or linguistics. After the initial training phase, most of the annotation work was carried out as single annotation. However, in order to monitor annotation consistency, double annotation batches were assigned regularly. In double annotation, one annotator first extracted the candidate paraphrases from the aligned documents, but later on these candidates were assigned to two different annotators, who annotated the labels independently from each other. Afterwards, the two individual annotations were merged and conflicting labels resolved together with the whole annotation team. These consensus annotations constitute a consolidated subset of the data, which can be used to evaluate the overall annotation quality by measuring individual annotators against this subset.

A total of 2025 examples (2% of the paraphrases in the corpus, excluding rewrites) were double annotated, most of these being annotated by exactly two annotators; however, some examples may include annotations from more than two annotators, and thus the total amount of individual annotations for which the consensus label exists is bit more than twice the number of double annotated examples (4287 annotations in total). We measure the agreement of individually annotated examples against the consolidated consensus annotations in terms of accuracy, that is the proportion of individually annotated examples where the label matches the consensus annotation.

The overall accuracy is 70% when using the full annotation scheme (base labels 1–4 as well as all flags). When discarding the least common flags *s* and *i* and evaluating only base labels and directional subsumption flags, the overall accuracy is 74%.

In addition to agreement accuracy, we calculate two versions of Cohen's kappa, a metric for inter-annotator agreement taking into account the possibility of agreement occurring by chance. First we measure the kappa agreement of all individual annotations against the consolidated consensus annotations, an approach typical in paraphrase literature. This kappa is 0.63, indicating substantial agreement. Additionally, we measure the Cohen's kappa between each pair of annotators. The weighted average kappa over all annotator pairs is 0.42 indicating moderate agreement. Both are measured on full labels. When evaluating only on base labels and directional subsumption flags, these kappa scores are 0.66 and 0.45, respectively.

Direct comparison of annotation agreement with other manually annotated paraphrase corpora is not straightforward due to several factors affecting the expected agreement measures, the most influential factors likely being the labeling scheme and label distribution of the corpora. While the kappa measure tries to account for this, this is especially true for accuracy. It should also be noted that in many semantic annotation tasks, agreement scores can only be used as estimates, and low score does not necessarily refer to a low annotation quality, but rather the nature of the task itself. (Pavlick and Kwiatkowski (2019), Davani, Díaz, and Prabhakaran (2022)) When comparing to other paraphrasing projects, all our metrics are in the same ballpark with other manually annotated samples, MRPC reporting accuracy of 84% with binary labels, Opusparcus accuracy between 64% and 67% with four labels, and ParaSCI reporting kappa of 0.71 when measuring the individual annotator against the majority vote on a five label scheme. Furthermore, one must also note that while our manual annotation primarily focuses on distinguishing between different positive labels, the other annotation efforts mentioned include also substantial amount of negatives, making the task slightly different from ours.

6.1.1. Rewrites

As mentioned earlier, during the annotation, the annotators have the possibility to rewrite the statements if the classification is anything else than pure label 4. This can be interpreted as the annotators fixing all flaws in the paraphrases and turning the candidates into perfect, context independent paraphrases. In order to evaluate the assumption of the rewrites always being a pure label 4, we sample 500 rewrites for double annotation. To ensure that the annotator does not know whether the candidate is a rewrite or normal extracted paraphrase, the rewrites are mixed together with normal paraphrase candidates in a 50/50 ratio. In addition, during this experiment, the document context is hidden in the annotation tool, as the context has a potential to reveal the candidate being a rewrite. The data are distributed in a fashion where all annotators receive only candidates previously annotated by someone else so that there is no risk of the annotators recalling the previously annotated examples. The candidates are also randomly shuffled.

After merging and resolving the double annotated examples, 78% of rewrites received the label 4. This is on par with the overall annotation consistency, showing the quality of rewrites largely following that of the natural examples in the corpus.

6.2. Lexical diversity and corpus comparison

One of our main goals was to obtain a set of paraphrase examples that are not highly lexically similar. In Figure 2, we measure the distribution of different labels in the corpus conditioned on the cosine similarity of the paraphrase pairs calculated using TF-IDF weighted character n-grams of lengths 2–4. While the different positive labels are evenly distributed in the low lexical similarity area up until similarity value 0.5, in the high similarity area the label 4 begins to dominate the data.

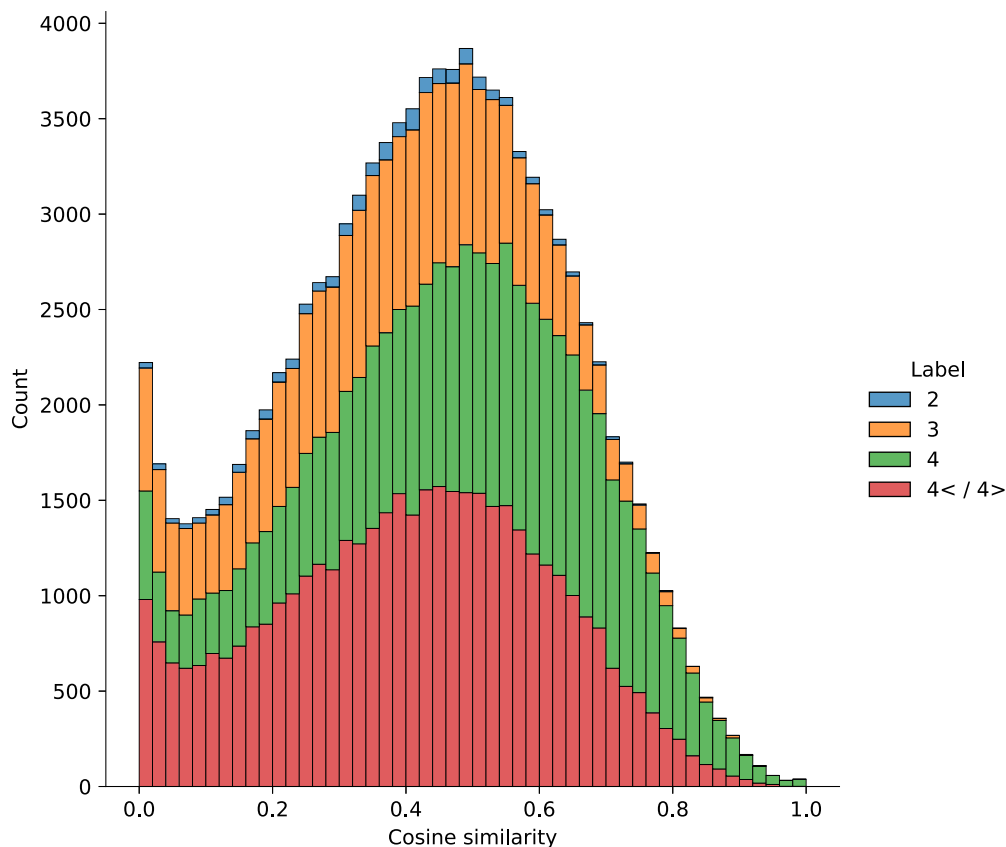


Figure 2. Histogram of different labels in the corpus conditioned on cosine similarity of the paraphrase pairs.

However, as can be seen from the figure, most of the paraphrases in the corpus fall into the low or mid-range similarity area making the high similarity quite sparsely populated.

Next, we compare our corpus with the two existing Finnish paraphrase candidate corpora, Opusparcus and TaPaCo using three different metrics: (1) the distribution of the lengths of the paraphrased segments, (2) the distribution of lexical similarity values of the two paraphrased statements, and (3) the presence of systematic paraphrasing patterns that can be identified automatically.

Such direct comparison between different corpora is naturally complicated by several factors. Firstly, compared with our manually annotated paraphrases with significant bias towards positive labels, both Opusparcus and TaPaCo consist primarily of automatically extracted paraphrase candidates, and the true label distributions are mostly unknown. The small manually annotated development and test sections of Opusparcus are sampled to emphasize lexically dissimilar pairs, and therefore not representative of the characteristics of the rest of the corpus, limiting their usage for corpus comparison purposes. We therefore compare with the fully automatically extracted sections of both Opusparcus and TaPaCo, as these represent the bulk of the corpora. In our corpus, we can discard the small proportion of examples of label 2, that is the examples known to not be paraphrases, while the automatically extracted sections of Opusparcus and TaPaCo are expected to include a significant portion of negative paraphrase examples as well. Therefore, when drawing any conclusions an important factor to consider is that the characteristics of false and true candidates may differ substantially, false candidates for example likely being on average more dissimilar in terms of lexical overlap than true candidates.

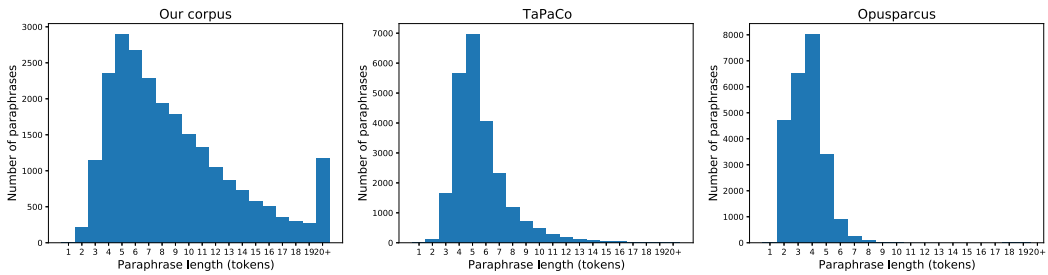


Figure 3. Comparison of paraphrase length distributions in terms of tokens per paraphrase.

For each corpus, we sample 12,000 paraphrase pairs in order to keep the sizes of the compared sets uniform. For our corpus, we selected a random sample of true paraphrases from the train section. For TaPaCo, the sample covers all paraphrase candidates from the corpus, however with the restriction of taking only one, random pair from each ‘set’ of paraphrases, while for Opusparcus, which is sorted by a confidence score in descending order, the sample was selected to contain the most confident 12K paraphrase candidates.^g

From Figure 3, it can be seen that the distribution of the paraphrase lengths in our corpus is wider and contains a hatriivial amount of longer paraphrases as well, while the other two corpora mainly contain relatively short paraphrase candidates. The average number of tokens in our corpus is 8.8 tokens per one paraphrase statement, while it is 5.6 in TaPaCo and 3.6 in Opusparcus. Furthermore, as the manual paraphrase extraction was not tied to follow sentence boundaries in our corpus, we measure how many of our paraphrases are short phrases, single sentences, or longer than a sentence. To this end, we apply a Finnish dependency parser (Kanerva *et al.* 2018) to segment sentence boundaries and recognize whether a sentence is well-formed (starts with a capitalized letter, ends with a punctuation character and includes a main verb) or not. We find that approximately 12% of the paraphrase statements are phrases or not well-formed single sentences, 73% are well-formed, single sentences, 13% are two sentences long, and the remaining 2% being segments which are more than two sentences long. When looking into paraphrase pairs instead of individual paraphrase statements, 63% of the pairs have one-to-one mapping of well-formed sentences, following with one-to-two (10%), sentence-to-phrase (9%), phrase-to-phrase (7%), and two-to-two (7%) mappings, the other variants occurring only rarely.

Figure 4, the cosine similarity distribution of the paraphrase pairs is measured using TF-IDF weighted character n-grams of length 2–4 for these three corpora. This allows us to establish to what degree the corpora contain highly lexically distinct pairs. From this figure, it can be seen that our corpus has a larger proportion of paraphrases with lower lexical similarity, while the distribution of the other two corpora are skewed towards pairs with higher lexical overlap.

Finally, we study the corpora from the point of view of systematic paraphrasing patterns, that is pairs which are formed in a systematic, predictable manner. To this end, we follow the method used in our prior work (Chang *et al.* 2021), recognizing six systematic ways in which the two segments of a paraphrase pair differ from each other: (1) word reordering, (2) word inflections (both having same lemmas in the same order), (3) lemma reordering, (4) lemma reordering after excluding all functional words (both having the same content word lemmas), (5) synonym replacements, and (6) a combination of (4) and (5).^h These six types of differences are automatically detectable

^gWhen the length analysis was repeated with a sample of 480K most confident pairs, the length distribution and average length remained largely unchanged, while the similarity distribution became close to flat. Without manual annotation, it is hard to tell the reason for this behavior.

^hIf a paraphrase pair can be accounted by either disregarding functional words or synonym substitution, it is classified as disregarding functional words.

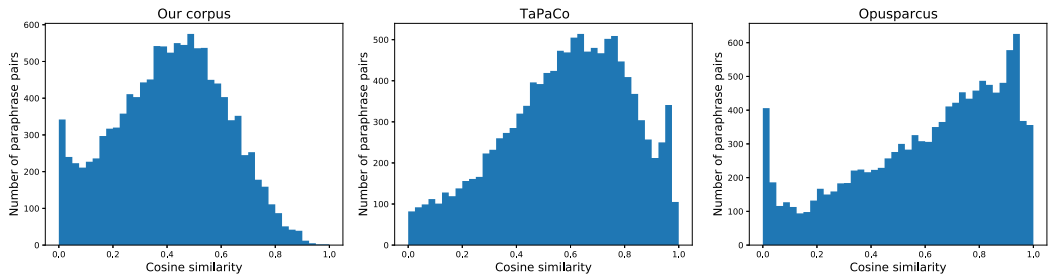


Figure 4. Comparison of paraphrase pair cosine similarity distributions.

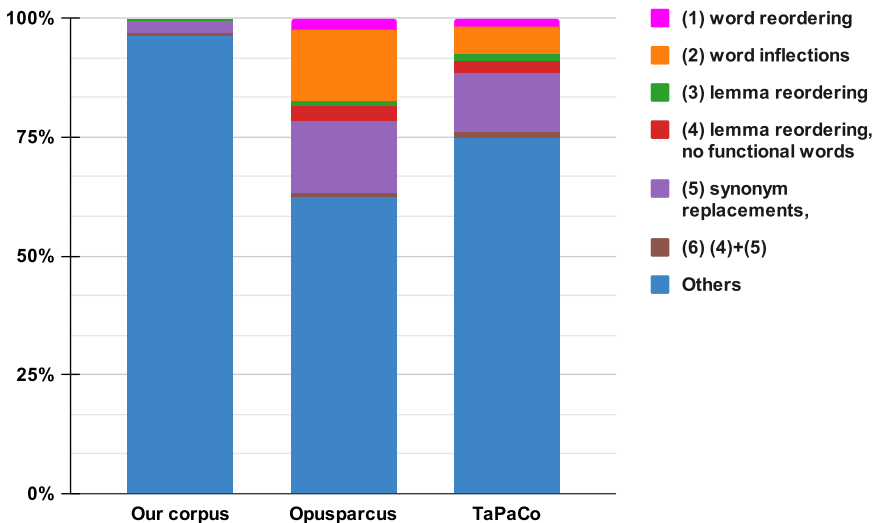


Figure 5. Percentage of the types of systematic differences characterizing the paraphrases in Opusparcus, TaPaCo, and our corpus. *Others* refers to all paraphrases including differences not automatically detectable by the used method.

with a simple approach and can be therefore regarded as to some degree “trivial” paraphrase pairs. From Figure 5, it can be seen that our corpus has a notably smaller proportion of trivial paraphrases than Opusparcus and TaPaCo. While the other two corpora have a larger proportion of paraphrases that can be accounted for by lemmatization, that is type (1), (2), and (3), our corpus has less than 1% of these each. The most prominent type of trivial paraphrases in our corpus is synonym replacement, at 2%. These results support that our manually extracted paraphrases contain more interesting, nontrivial paraphrases than automatically collected corpora and help to validate our manual extraction approach.

7. Paraphrase classification

Having described the paraphrase corpus itself, we will continue to paraphrase modeling experiments. We first apply a pairwise paraphrase classifier, where for a given candidate pair the classifier predicts the label based on the labeling scheme used in the corpus. While the classification model could be straightforwardly trained using only the annotated paraphrase corpus, in addition to such a baseline model we also apply a bootstrapping approach where the training data is augmented with automatically extracted negative pairs to account for the low frequency of negative pairs in the original corpus.

When creating the paraphrase corpus, we concentrated on building a dataset of nontrivial paraphrases classified as positive in manual annotation (label 3 and above), where the occasional label 2 paraphrase candidates were only a by-product of the annotation work. However, in order to train models able to distinguish negative candidates from the positives, a sufficient number of negative examples is required during the model training. While unrelated negative candidates (label 1) can be obtained trivially by pairing arbitrary sentences, it is shown for example by Guo *et al.* (2018) in the context of parallel data mining that it is not sufficient to introduce negatives based only on arbitrary pairs. Instead, better results can be obtained by including hard negatives, that is candidates which share for example topic or are otherwise related while still not being paraphrases.

In order to obtain such training data for the paraphrase classifier, in our bootstrapping approach we use sentence embeddings obtained from a basic language model without task-specific fine-tuning to select semantically related pairs of sentences from a large corpus of text. These are subsequently filtered using an initial classifier trained purely on the manually annotated corpus data, preserving examples with a confident negative prediction. Finally, we train new models for paraphrase classification using a combination of the manually annotated corpus and the automatically extracted negative candidates. Next, we describe all these steps in detail.

7.1. Paraphrase classifier

Our paraphrase classification model is a pairwise classifier based on the BERT encoder, following our initial work reported in Kanerva *et al.* (2021b). The model receives one candidate pair at a time, encoded as the sequence [CLS] A [SEP] B [SEP], where A and B are the two paraphrase statements and [CLS] and [SEP] the special tokens in the BERT model. The classifier is a multi-output model implemented on top of the pretrained FinBERT language model (Virtanen *et al.* 2019), including four separate prediction layers, one for the base label (with classes 2, 3, or 4), one for the subsumption flag (<, > or none), one for the style flag (s or none), and one for the minor deviation flag (i or none). As the additional flags only apply to examples where the base label is 4, no gradients are produced for subsumption, style, and minor deviation prediction layers if the base label of the example is 2 or 3. The predictions are based on five different embeddings obtained from the final BERT layer: embeddings for the [CLS] and the two [SEP] tokens, as well as the average of token embeddings calculated separately for statement A and statement B, all five concatenated together and projected for the four prediction layers. The overall model design (e.g. concatenating the five embeddings rather than using the plain [CLS] embedding) is optimized during preliminary experiments conducted on the development data. The use of multiple output layers rather than treating each label combination a separate class in standard multiclass classification is chosen to account for certain flag combinations, such as 4>is, which would not be predicted at all by a standard multiclass model as such label combinations are so rare in the data.

The initial classifier is trained on the Turku Paraphrase Corpus using the data split reported in Table 3, receiving an accuracy of 58.1 and a weighted average F-score of 57.6 when tested on the corpus test set treating each complete label as its own class during evaluation. As expected, the initial classifier is weakest at classifying the small amount of negative examples (label 2) in the test set, giving an F-score of 30.3 for label 2, and fully reflecting the design choices of the corpus. The full evaluation numbers for the initial classifier are given later in Section 7.3 (Table 4) where the results are compared with the final, bootstrapped model.

7.2. Extracting candidate pairs for model bootstrapping

Deep language models, such as BERT (Devlin *et al.* 2019) or LASER (Schwenk and Douze 2017), are commonly used as general sentence encoding methods, assigning dense vector representations to sentences and other short text segments. Simple metrics, such as cosine similarity or Euclidean distance, can then be used to efficiently estimate the similarity of two sentences in the vector space,

Table 4. Baseline classification performance on the two test sets, when the base label and the flags are predicted separately. In the upper section, we merge the subsumption flags with the base class prediction, but leave the flags *i* and *s* separated. The rows *W. avg* and *Acc* on the other hand refer to performance on the complete labels, comprising all allowed combinations of base label and flags. *W. avg* is the average of P/R/F values across the classes, weighted by class support. *Acc* is the accuracy

Turku Paraphrase Corpus test set					Opus-parsebank-test				
Label	Prec	Rec	F	Support	Label	Prec	Rec	F	Support
2	46.8	22.4	30.3	161	neg	99.0	23.1	37.5	6712
3	60.3	50.9	55.3	2434	3	11.7	48.3	18.8	1146
4<	55.8	57.9	56.8	2003	4<	36.9	64.7	47.0	425
4>	57.0	61.9	59.4	2287	4>	37.8	70.7	49.3	560
4	70.5	74.3	72.4	3586	4	47.1	91.3	62.1	793
<i>i</i>	50.0	47.4	48.6	454	<i>i</i>	52.0	71.3	60.2	164
<i>s</i>	49.1	37.0	42.2	438	<i>s</i>	28.2	48.0	35.6	50
<i>W. avg</i>	57.7	58.1	57.6		<i>W. avg</i>	77.8	35.5	37.9	
<i>Acc</i>			58.1		<i>Acc</i>			35.5	

with sentences equivalent or closely related in meaning being also highly similar in terms of these metrics. We rely on such embedding similarities in order to find promising, initial candidates of related sentences for model bootstrapping, where our aim is to collect negative pairs including a nontrivial topical overlap (hard negatives). For creating the sentence embeddings, we use the vanilla BERT model pretrained for Finnish without any task specific fine-tuning of the model.

In order to obtain enough candidate sentences for collecting hard negatives for our bootstrapping experiments, we use two different data sources: OPUS and Finnish Internet Parsebank. OPUS (Tiedemann 2012) is an open parallel corpus collecting a diverse set of parallel sentences ranging from EU legislation and software manuals to movie subtitles. The OPUS data is obtained through the data release of the Tatoeba translation challenge (Tiedemann 2020). The Finnish Internet Parsebank (Luotolahti *et al.* 2015) is a large-scale Finnish corpus collected through dedicated web crawls targeted to find high quality Finnish material from the Internet. Together, these two resources include almost 400M unique sentences. All unique sentences in this collection are first encoded with the FinBERT model of Virtanen *et al.* (2019) taking the average of token embeddings to obtain one vector for each sentence. Next, for each sentence, its five most similar sentences are collected from the same source (OPUS or Parsebank) using Euclidean distance of the embeddings implemented in the FAISS library (Johnson, Douze, and Jégou 2021) for fast similarity comparison, constituting a massive candidate set of $400M \times 5$ closely related sentence pairs. Finally, all duplicate pairs (irrespective of direction) are discarded.

To understand the distribution of different paraphrase labels in this set of candidates, we selected a random sample for manual annotation. A total of 15,000 sentence pairs are sampled, taking 7500 pairs from both OPUS and Finnish Internet Parsebank. So as to maximize the informativeness of this manual evaluation, we stratify the sample in terms of lexical similarity, measured as cosine similarity of term frequency (TF) vectors based on character *n*-grams of lengths 2–4. All candidate pairs are split into 20 lexical similarity intervals in increments of 0.05, with an equal number of pairs selected from each interval for manual annotation. This stratified sampling together with manual annotation allows us to estimate the distribution of different labels in each similarity interval separately. In the manual annotation, we labeled 14,530 candidate pairs (470 were skipped with label *x* during the annotation due to various issues such as incorrect

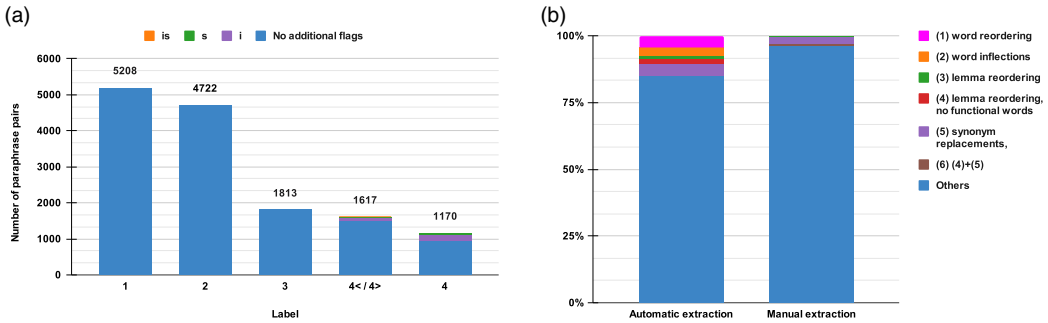


Figure 6. (a) Distribution of the manually annotated labels in the opus-parsebank set including both development and test examples. (b) Comparison of the types of paraphrases in the manually and automatically extracted data. The *manually-extracted data* refers to the training set of our corpus, while the *automatically extracted data* refers to the combination of opus-parsebank-dev and opus-parsebank-test sets.

language or whitespace-only differences). The sample is divided into development and test sections (hereafter opus-parsebank-dev and opus-parsebank-test), with a 1/3 and 2/3 split. While the development section is used to analyze the different properties of the data, the test section is reserved only for the final test purposes, and none of the annotated data is used for the actual model training.

Next, we analyze the annotated sample from several perspectives. In Figure 6, on the left we show the label distribution of this sample, and on the right side we plot the automatically detectable systematic paraphrasing patterns introduced in Section 6.2. Contrary to the manually extracted corpus, the sample does not strive to exclude uninteresting candidates including only elementary variation, and among the examples with a high lexical similarity, trivial differences are included, such as differences purely in punctuation or capitalization. While the manually constructed corpus included only occasional negative paraphrases, as expected, the label distribution in the opus-parsebank sample is skewed towards negative paraphrases (68% being annotated with label 1 or 2). When measuring the automatically detectable systematic paraphrasing patterns among positive examples (labels 3 and above), the figure confirms the higher tendency towards trivial variation appearing among the automatically extracted paraphrases than among the manually selected ones. Along with those shown in the figure, additional 2% of positive paraphrase pairs in the opus-parsebank sample contain only small character differences that are usually typos or punctuation differences, totaling the recognized elementary variation to cover approximately 17% of all positive paraphrase pairs. However, part of the elementary variation can be explained by the stratified sampling over lexical similarity values, as high similarity areas have proportionally more elementary variations, compared with the automatically detected paraphrase candidates with lower similarity ranges, which are mostly negative paraphrase pairs, along with some nonelementary positive paraphrase pairs.

Finally, we analyze the annotated sample regarding the reliability of the classifier prediction scores, with the aim of identifying areas where we can be reasonably confident in the classifier predictions and sample “safe” negative examples to complement the primary manually annotated corpus. When simultaneously plotting classifier prediction scores (probability of negative label) together with the lexical similarity intervals into a two-dimensional plot, we are able to divide the examples into several tiles, which can furthermore be enriched with the manually annotated labels to estimate the actual amount of negative candidates (labels 1 and 2) in each tile. This information can be used to select tiles (and their corresponding lexical similarity and prediction score values) to collect safely negative or safely positive paraphrase candidates when applying the same metrics for the full collection of closely related pairs. The observed tiles are demonstrated in Figure 7,

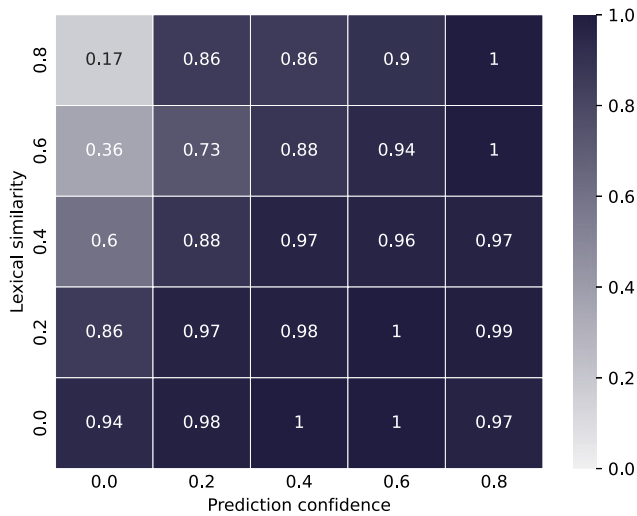


Figure 7. Heatmap with estimated negative example density per tile in increments of 0.2 for opus-parsebank-dev. Lexical similarity is plotted in y-axis and prediction confidence in x-axis, creating two-dimensional tiles when both are divided in increments of 0.2. Each tile is yet enhanced with a density score indicating the percentage of negative examples in the tile based on the manually annotated labels.

where the data is split into five intervals in increments of 0.2 on both axes, as both the prediction score and lexical similarity values range between 0 (unsure, highly dissimilar) and 1 (confident, highly similar). Each tile is yet enhanced with an annotation indicating the percentage of negative labels in the tile estimated using the manually annotated sample.

For collecting negative candidates, all pairs with lexical similarity of under 0.1 or negative class prediction confidence over 0.4 were chosen as optimal region. When applying these values across the whole set of closely related sentence pairs (discarding those in the annotated sample), we were able to extract approximately 5M nonparaphrase candidates with precision of 97.7% as estimated from the manually annotated sample. Additionally, the same experiment was repeated for the positive paraphrase candidates by using lexical similarity of over 0.5 and the model's prediction confidence score of 0.998 or greater for the base label 4, obtaining a set of 500K positive paraphrase candidates with estimated precision of 95.8%. Both datasets are released as supplementary data together with the manually annotated examples in order to support for example training with a binary objective (paraphrase or not-a-paraphrase).

7.3. Paraphrase classification results

For the final classification experiments, the manually annotated Turku Paraphrase Corpus training set of 84K pairs is combined with an additional 84K pairs sampled from the automatically gathered negatives, therefore creating a somewhat balanced set of positive and negative training examples. While all manually annotated examples naturally include the full label information, for automatically gathered “training” negatives, we do not have distinction between the two negative labels (label 1 and label 2), and therefore we opted to use only a single label for all negative examples while training the classifier. As shown in Table 5 versus the baseline performance shown in Table 4, besides slightly improving the label 2 classification performance, enhancing the training data with automatically gathered negatives does not affect the performance on the Turku Paraphrase Corpus test set, where the great majority of the test examples fall into the different positive labels. Therefore, the automatically gathered negative training examples do not

Table 5. Final classification performance on the two test sets, as in Table 4

Turku Paraphrase Corpus test set					Opus-parsebank-test				
Label	Prec	Rec	F	Support	Label	Prec	Rec	F	Support
2	40.2	32.9	36.2	161	neg	95.0	75.0	83.8	6712
3	59.3	52.6	55.8	2434	3	25.2	36.3	29.8	1146
4<	56.0	58.1	57.0	2003	4<	44.7	62.4	52.1	425
4>	58.3	59.8	59.1	2287	4>	46.0	68.0	54.9	560
4	70.5	73.9	72.2	3586	4	56.3	89.7	69.2	793
i	51.8	48.9	50.3	454	i	56.0	71.3	62.7	164
s	49.4	37.7	42.8	438	s	32.0	48.0	38.4	50
W. avg	57.9	58.3	58.0		W. avg	78.1	69.9	72.6	
Acc			58.3		Acc			69.9	

seem to decrease the performance of positive predictions. However, in the opus-parsebank-test, where more than two-thirds of the examples are negatives and therefore larger differences can be expected, the bootstrapped model significantly outperforms the baseline model on the negative class, increasing the negative class F-score from 37.5 to 83.8, which is mostly caused by heavily increasing its recall without compromising the precision too much. This naturally also increases the precision of the positive classes by not as heavily overpredicting the positives; however, the classifier still struggles in distinguishing between different positive labels, as well as precisely setting the border between negatives and contextual paraphrases. When compared with the estimated human performance on the task, the classifier is still almost 12pp behind the accuracy of the human annotators when measured on the Turku Paraphrase Corpus test set. However, in contrast to the humans, the current model does not have access to the document context, which may naturally complicate the labeling decision particularly in the context dependent cases (label 3). In the future, we plan to extend the classification work towards context-aware models.

8. Fine-tuned sentence embeddings in paraphrase mining

Paraphrase classification has been shown to work well and is expected to give good results accuracy-wise when judging the paraphrasability of a candidate pair of statements. However, the pair-wise classification approach becomes infeasible especially in large-scale paraphrase retrieval applications, as it requires applying the computationally heavy classifier separately for each possible candidate pair. In large-scale scenarios such as paraphrase mining where the objective is to find good paraphrase candidates from a large collection of sentences, the number of candidate pairs is quadratic. Therefore, computationally a much more feasible approach is to pre-compute sentence embeddings once, and for each candidate pair apply only a computationally light-weight metric (e.g. cosine similarity or Euclidean distance) using these pre-calculated representations. In addition to directly applying a pre-trained language model such as BERT, one can also optimize the representations for paraphrase comparison by fine-tuning these models to create sentence embeddings such that paraphrased statements receive a high similarity score when comparing the calculated embeddings using for example cosine similarity, while semantically unrelated statements receive a low similarity score.

A well-known model of this kind is the Sentence-BERT (SBERT) (Reimers and Gurevych 2019), where the training objective is to improve individual sentence embeddings in order to better support their direct cosine similarity comparison. The SBERT fine-tuning objective applies a siamese network encoding, where the two sentences are first encoded individually producing a fixed size embedding for both, and these embeddings are then fine-tuned through either a classification or cosine similarity objective. The SBERT models are typically trained on semantically related sentences taken from corpora gathered for for example paraphrasing, natural language inference or translation, where the positive pairs are mixed with unrelated sentence pairs in order to provide also negative training examples.

Next, we train a Finnish SBERT model for the paraphrasing task and evaluate it on the task of paraphrase retrieval using the corpus data. In addition to the paraphrase corpus, we evaluate the fine-tuned embedding model also in a large-scale paraphrase mining experiment using a dataset of almost 400M candidate sentences.

8.1. SBERT training and evaluation

In the following, we evaluate the SBERT sentence embedding model on our corpus in the context of paraphrase mining. We train a Finnish SBERT model initialized from the pre-existing Finnish BERT-base model with our paraphrase data. We use batch size of 16 and mean pooling over the final BERT layer, the best-performing pooling method in the original SBERT work (Reimers and Gurevych 2019). Since the goal is to identify paraphrase candidates, we collapse the labels into binary: labels 1 and 2 becomes negative, and labels 3 and above positive. We experiment with different combinations of training datasets: (1) the manually annotated Turku Paraphrase Corpus training set (`train`), consisting of 81.8K positive and 1.4K negative pairs, (2) the manually annotated training set and the full set of automatically gathered negatives (`train+neg`), with 81.8K positives and 5.6M negatives, (3) the manually annotated training set and the full sets of automatically gathered positives and negatives (`train+neg+pos`), totaling 625K positives and 5.6M negatives, and (4) only the automatically gathered positives and negatives (`neg+pos`), with 543K positives and 5.6M negatives. The learning rate is optimized on the development set, using the value $1e-5$ for all four experiments.

As a non-neural baseline, we use TF-IDF representations of character bi- and tri-grams. As a modern, neural baseline, we use the vanilla Finnish BERT model to directly encode single sentences without any task specific fine-tuning. For hyperparameter optimization, we test CLS vector, mean-pooling, and max-pooling on the development set, and select mean-pooling as the final pooling method.

We evaluate these models on the paraphrase retrieval task, that is given the statement s_1 from a known paraphrase pair (s_1, s_2), how well the model is able to identify its corresponding paraphrased version s_2 from a collection of Finnish sentences using cosine similarity. First, we evaluate the retrieval among all paraphrase statements in the corresponding manually annotated test sets. That is, we take both statements from all paraphrase pairs in the corpus test set and deduplicate them. This gives 19,893 unique statements in the Turku Paraphrase Corpus test set and 19,271 in the opus-parsebank-test set. All these candidate text segments are first embedded, and separately for each paraphrase statement in the test set, the candidates are sorted in descending order based on cosine similarity giving the most similar candidates first. A good embedding model is expected to give higher cosine similarity for a paraphrase pair than for a random segment pair, that is rank the known paraphrase pair high in the sorted candidates.

First we measure top-1 retrieval accuracy of all positive examples (labels 3 and above). This is to inspect how likely the model ranks a good paraphrase pair first among candidate sentences if the corresponding paraphrased version is guaranteed to exist in the collection. The results are given in Figure 8. When measured on the Turku Paraphrase Corpus test set (blue color in the figure),

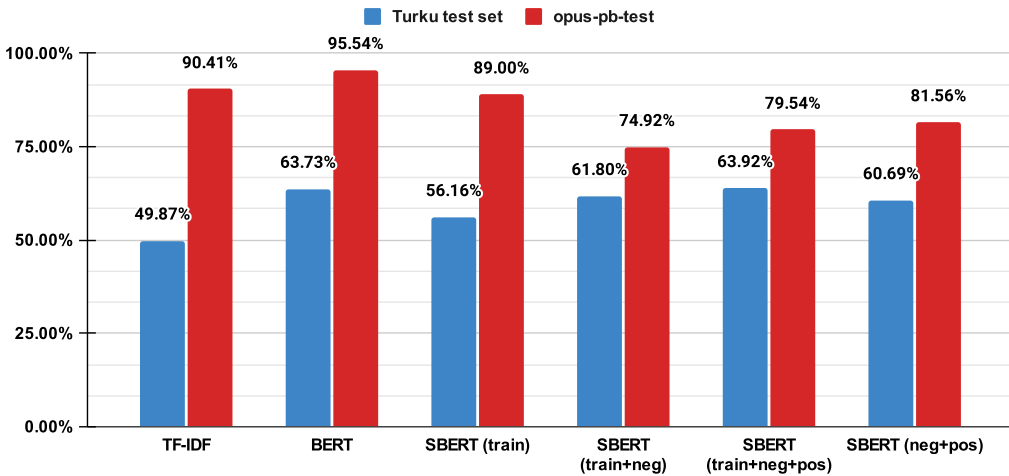


Figure 8. The top-1 retrieval accuracy (higher is better) of all positive paraphrases in the Turku Paraphrase Corpus test set and the opus-parsebank-test set. The test sets consists of 19,893 and 19,271 unique retrieval candidates respectively. The exact accuracy numbers are visualized on top of the bars.

the SBERT model *train+neg+pos* gives comparable, if not slightly better, results to the vanilla BERT baseline. The other SBERT models underperform vanilla BERT in terms of top-1 accuracy. Unsurprisingly, all the neural models outperform the TF-IDF method. When considering the opus-parsebank-test set, where the paraphrase candidates were sampled based on a combination of the TF-IDF and FinBERT similarity scores, it is not a surprise to see that these two methods obtain the highest performance. While all examples in the opus-parsebank-test set are selected based on their high BERT similarity score, fully explaining the high top-1 accuracy of the BERT model, the sample was stratified to include examples from all lexical similarity areas. However, after manual annotation most of the positive examples are actually located in the high similarity area (the average lexical similarity of positive examples being 0.73 compared with 0.5 on the full development sample), therefore to some extent skewing the evaluation also in terms of TF-IDF similarity.

However, measuring the top-1 accuracy of the positive paraphrases does not take into consideration how these models perform on the negative pairs, where the model should not give a high similarity for nonparaphrase pairs even if their lexical similarity is high. In the light of this, we next measure the average ranking positions of the paraphrase candidates separately for each label in order to see whether fine-tuning the model successfully decreases the similarity of negative paraphrase pairs while increasing or maintaining the similarity of the positive pairs, as it is expected that a good model should give lower similarity and therefore also worse ranking positions for unrelated candidates than for related candidates, while similarity of related candidates in turn should be lower than similarity of real paraphrases and so on. To measure this effect, in Figures 9 and 10, we report average ranking positions in percentage for each label separately, where the actual on average ranking positions are normalized to percentages, so as there were 100 candidates. This means that a perfect ranking, where the correct candidate is always ranked top-1, would give 0%, whereas the results of 5% means that the correct candidate is on average ranked 6th out of 100 candidates.

Based on the results in Figure 9, our ranking assumption seems to hold in the sense that the more universal the paraphrase pair is the better the average ranking position seems to be in general. However, with the exception of the vanilla BERT and SBERT trained on the Turku Paraphrase

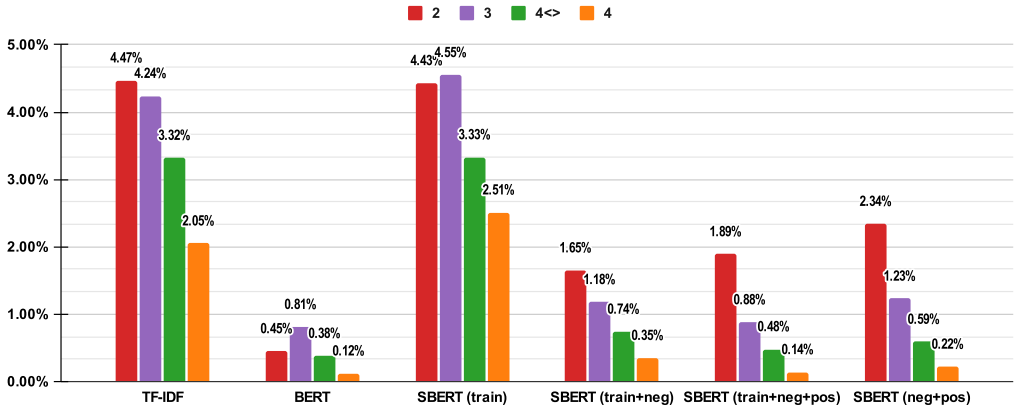


Figure 9. The average ranking positions normalized to percentages (lower is better) for the Turku Paraphrase Corpus test set by various models. The ranking is measured separately for each paraphrase label (2, 3, 4</>, and 4), however disregarding the flags i and s. The exact numbers are visualized on top of the bars (percentage calculated out of 19,893 candidate sentences).

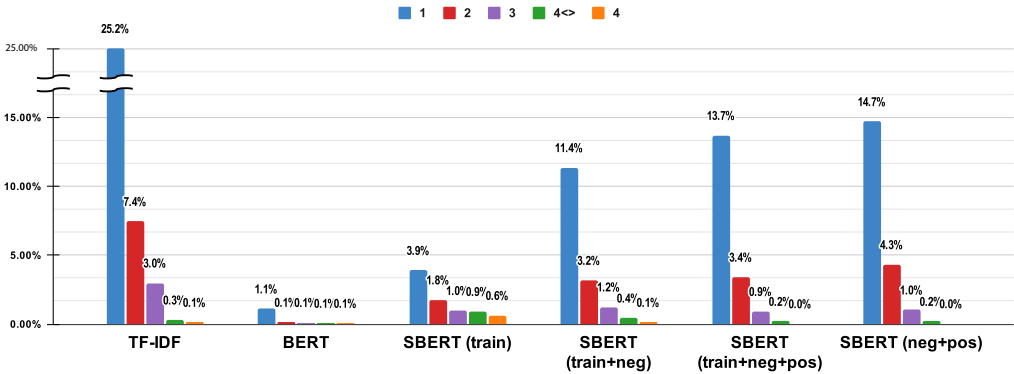


Figure 10. The retrieval of the opus-parsebank test set paraphrase candidates by various models. The numbers on top of the bars indicate the average ranking in percentage (out of 19,271 candidate sentences) for each class of paraphrase candidates. The ranking is measured separately for each paraphrase label (1, 2, 3, 4</>, and 4), however disregarding the flags i and s.

Corpus only (train, where fine-tuning data does not include practically at all negatives) models do not distinguish between the negative label 2 and positive label 3, therefore giving high similarity scores also for negative paraphrase pairs. However, when increasing the amount of negative examples seen during the training, the fine-tuned SBERT models start to give clearly worse ranking positions for label 2 pairs compared with label 3 pairs as the model learns to judge these as negative examples, which appears to be the main advantage of SBERT models over the vanilla BERT. When comparing the different SBERT models, the observations remain largely the same as in the top-1 accuracy analysis. That is, the SBERT model trained with all available training data yielding the best results among the fine-tuned models. For the evaluation on the opus-parsebank-test set (Figure 10), the average of BERT embeddings clearly achieves the best ranking positions, which is not at all surprising as the test set was selected based on the similarity of BERT embeddings. Again, the notable fact is that while the original BERT naturally assigns good ranking positions for the negative examples in this dataset as well (label 1 and 2), the model fine-tuning clearly helps to distinguish between positive and negative examples, pushing the negative examples further while only negligibly affecting the ranking for the positive examples.

8.2. Large-scale paraphrase mining

A larger collection of Finnish candidate sentences presumably makes the paraphrase mining task more difficult as the number of difficult distractors also increases. For instance, considering top-1 accuracy, it takes only one incorrect distractor sentence to fool the model. Thus, we simulate a realistic paraphrase mining setting by mining the correct target sentence among the combined set of 399M unique sentences from the combination of the Finnish Internet Parsebank, OPUS, and our paraphrase corpus. First, we calculate and index the SBERT embedding for each sentence in this large combined dataset. Then, for each test set paraphrase pair (s_1, s_2), we query the index with the embedding of s_1 and measure at which rank out of the nearly 400M candidates the embedding of s_2 is found in terms of Euclidean distance. For comparison, we also carry out the same experiment with the vanilla FinBERT model embeddings, so as to establish whether the fine-tuning of the SBERT model translates into better performance on the sentence similarity task, as well as with the multilingual SBERT model `paraphrase-xlm-r-multilingual-v1` released by Reimers and Gurevych (2019) fine-tuned to create comparable embeddings for over 50 languages. The multilingual SBERT model is based on the monolingual English SBERT trained on a massive collection of semantically similar English sentence pairs, and the multilingual XLM-RoBERTa-base language model (Conneau *et al.* 2020), where the multilingual language model was fine-tuned to mimic the embeddings of the English SBERT using multilingual knowledge distillation (teacher–student framework) on parallel data for over 50 languages.

The results are summarized in Figure 11, where we report the top-N accuracy (where $N=1, 10, 100, 1000, \text{ and } 2048$, which is the upper technical limit in the experiment) for label 3, label 4> or 4<, and label 4 separately. Most importantly, for the Finnish SBERT model (named `sbert` in the figure), we can see that 53% of label 4 paraphrases, 41% of label 4> or 4< paraphrases, and 29% of label 3 paraphrases are ranked among the top 10 most similar sentences from the group of nearly 400M candidates. This demonstrates that the SBERT model is highly efficient at finding paraphrase pairs also in cases where the number of candidates is in the hundreds of millions. This opens the possibility for further paraphrase mining from even very large text collections. While it is obviously infeasible to apply an expensive pairwise classification model to all sentence pairs (in our case that would be on the order of 400M squared), one can use SBERT as an initial filter and then apply the pairwise classification model to the comparatively small number of top candidates (in our case 400M times 10 pairs if using the cut-off of top-10 candidates). Finally, as seen in Figure 11, the vanilla FinBERT (`bert` in the figure) not fine-tuned for the semantic similarity task produces notably worse results compared with the SBERT models, while both Finnish and multilingual (`xmlrsbert` in the figure) SBERT produce comparable results. This seems to indicate that the advantage of model fine-tuning starts to pay off when the number of candidates for the retrieval is substantially increased. With this massive candidate set, the SBERT models are likely better at filtering out topically and lexically difficult distractors, which did not show up when using a smaller candidate set. The implementation of this experiment was carried out using the FAISS library (Johnson *et al.* 2021) for efficient GPU-accelerated k-nearest-neighbor vector similarity search in large vector collections.

9. Conclusions

In this paper, we presented the Turku Paraphrase Corpus, the first large-scale manually annotated corpus of Finnish paraphrases. The corpus contains 104,645 paraphrase pairs, targeted to create a challenging paraphrasing dataset suitable to test the capabilities of natural language understanding models. Each pair is manually labeled using a detailed annotation scheme. In addition to separating positive and negative paraphrase pairs, the annotation also distinguishes between paraphrases in all imaginable contexts and paraphrases in the given context but not necessarily elsewhere.

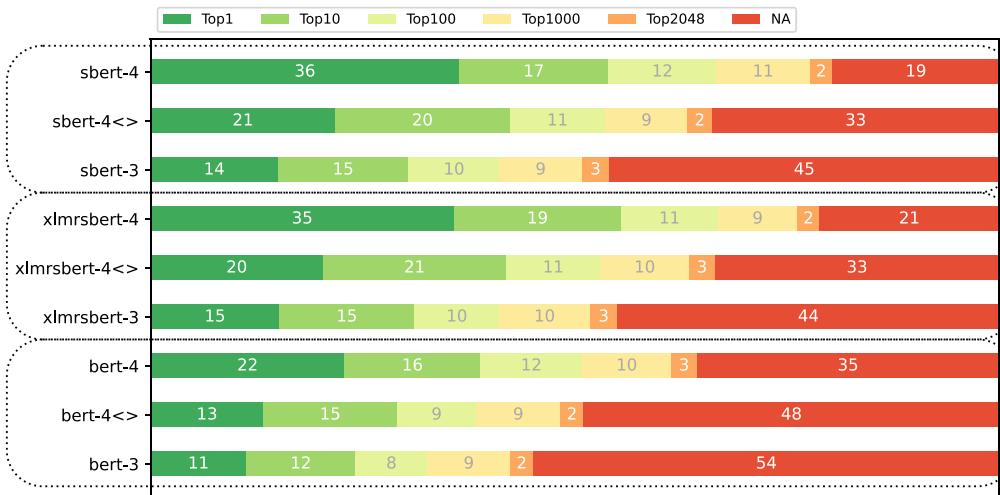


Figure 11. The retrieval of test set paraphrase pairs by the fine-tuned Finnish SBERT, the multilingual SBERT, and the vanilla FinBERT, out of 400M candidate sentences. The white numbers indicate percentage of pairs in the given category, and the retrieval is measured for the three main classes of paraphrase: 4, 4< or 4>, and 3 (disregarding flags s and i); and for several top k cut-offs. NA means that the correct sentence did not rank in the top 2048 list, which was the upper technical limit in the experiment.

The paraphrase pairs in the corpus are collected using a novel method for manual paraphrase candidate extraction, assuring both quality and variability of the extracted paraphrases, as well as efficiency in terms of person-months used for annotation. The paraphrases are manually selected from two related source documents, where a high tendency of naturally occurring paraphrases is expected. Compared with other paraphrase resources, the manual extraction is shown to produce notably longer and less lexically overlapping pairs than what automated candidate selection permits, creating a challenging dataset to be used for instance in evaluation of different language understanding models. In addition to quality, the advantage of manual candidate extraction is the possibility to collect and evaluate the paraphrase candidates in their original document context, setting many new possibilities for contextual paraphrase recognition. To our knowledge, this work is the first large-scale paraphrase corpus providing original document context information for the paraphrase pairs.

While 98% of the paraphrases in the corpus are manually classified to be at least paraphrases in their given context if not in all contexts (positive examples), in order to better facilitate also binary classification experiments (paraphrase or not-a-paraphrase), a method for semi-automatically extracting negative paraphrase candidates is presented, and a supplementary set of over 5 million negative paraphrase candidates is provided together with the actual corpus.

The initial modeling results confirmed the challenging nature of the dataset, giving weighted mean F-score of 58% for a pairwise classifier over the detailed annotation labels, the classifier accuracy substantially lacking behind the estimated human performance on the task. However, when applying semantic similarity models fine-tuned on the data for large-scale paraphrase mining from a collection of almost 400M candidates, the results were highly encouraging, the paraphrase retrieval model being able to rank the correct paraphrase pair among the top-10 for 29–53% of the evaluation examples depending on the paraphrase type.

While our initial paraphrase retrieval experiments show promising results, the classification experiments using the detailed labeling scheme are still far from human performance, indicating that the corpus can serve as a challenging evaluation task for different language understanding models. Such datasets have recently shown their importance when yet more powerful language

understanding models are approaching human-level performance on several popular evaluation sets, and more challenging tasks are introduced (Wang *et al.* 2019). However, despite our initial modeling experiments, there are still many new aspects to study with the dataset, such as how to utilize the contextual information available for the paraphrase pairs, and in the future work, we plan to further study the contextuality aspect of this data.

The corpus is available at github.com/TurkuNLP/Turku-paraphrase-corpus as well as through the popular HuggingFace datasets under the CC-BY-SA license.

Acknowledgements. We warmly thank Leena Salmi, Eriikka Paavilainen–Mäntymäki, Riikka Harikkala–Laihin and Veronika Laippala for their support in student data collection, as well as all anonymous students for agreeing to share their data. We gratefully acknowledge the support of European Language Grid which funded the annotation work. Computational resources were provided by CSC—the Finnish IT Center for Science and the research was supported by the Academy of Finland and the Digicampus project. We also thank Sampo Pyysalo for fruitful discussions and feedback throughout the project and Jörg Tiedemann for his generous assistance with the OpenSubtitles data.

Conflicts of interest. The authors declare none.

References

- Altheneyan A.S. and Menai M.E.B.** (2019). Evaluation of state-of-the-art paraphrase identification and its application to automatic plagiarism detection. *International Journal of Pattern Recognition and Artificial Intelligence* 34(4). <https://doi.org/10.1142/S0218001420530043>
- Arwinder Singh G.S.J.** (2020). Construction of paraphrasing dataset for Punjabi: A deep learning approach. *International Journal of Advanced Science and Technology* 29(06), 9433–9442.
- Bhagat R. and Hovy E.** (2013). Squibs: What is a paraphrase? *Computational Linguistics* 39(3), 463–472.
- Chang L.-H., Pyysalo S., Kanerva J. and Ginter F.** (2021). Quantitative evaluation of alternative translations in a corpus of highly dissimilar Finnish paraphrases. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age* online. Association for Computational Linguistics, pp. 100–107.
- Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V.** (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 8440–8451.
- Creutz M.** (2018). Open Subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA), pp. 1364–1369.
- Davani A.M., Daz M. and Prabhakaran V.** (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10, 92–110.
- Devlin J., Chang M.-W., Lee K. and Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186.
- Dolan W.B. and Brockett C.** (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, Korea. Asian Federation of Natural Language Processing, pp. 9–16.
- Dong Q., Wan X. and Cao Y.** (2021). ParaSCI: A large scientific paraphrase dataset for longer paraphrase generation. arXiv preprint arXiv:2101.08382.
- Eyecioglu A. and Keller B.** (2018). Constructing a Turkish corpus for paraphrase identification and semantic similarity. In **Gelbukh A.** (ed), *Computational Linguistics and Intelligent Text Processing*, Cham. Springer International Publishing, pp. 588–599.
- Federmann C., Elachgar O. and Quirk C.** (2019). Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China. Association for Computational Linguistics, pp. 17–26.
- Ganitkevitch J. and Callison-Burch C.** (2014). The multilingual paraphrase database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA), pp. 4276–4283.
- Ganitkevitch J., Van Durme B. and Callison-Burch C.** (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia. Association for Computational Linguistics, pp. 758–764.

- Gudkov V., Mitrofanova O. and Filippkikh E.** (2020). Automatically ranked Russian paraphrase corpus for text generation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, Online. Association for Computational Linguistics, pp. 54–59.
- Guo M., Shen Q., Yang Y., Ge H., Cer D., Hernandez Abrego G., Stevens K., Constant N., Sung Y.-H., Strope B. and Kurzweil R.** (2018). Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics, pp. 165–176.
- He Y., Wang Z., Zhang Y., Huang R. and Caverlee J.** (2020). PARADE: A new dataset for paraphrase identification requiring computer science domain knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 7572–7582.
- Johnson J., Douze M. and Jégou H.** (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 535–547.
- Kanerva J., Ginter F., Chang L.-H., Rastas I., Skantsi V., Kilpeläinen J., Kupari H.-M., Piirto A., Saarni J., Sevón M. and Tarkka O.** (2021a). Annotation guidelines for the Turku Paraphrase Corpus. Technical report, University of Turku, arXiv:2108.07499.
- Kanerva J., Ginter F., Chang L.-H., Rastas I., Skantsi V., Kilpeläinen J., Kupari H.-M., Saarni J., Sevón M. and Tarkka O.** (2021b). Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden, pp. 288–298.
- Kanerva J., Ginter F., Miekka N., Leino A. and Salakoski T.** (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*, Brussels, Belgium. Association for Computational Linguistics, pp. 133–142.
- Lan W., Qiu S., He H. and Xu W.** (2017). A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 1224–1234.
- Luotolahti J., Kanerva J., Laippala V., Pyysalo S. and Ginter F.** (2015). Towards universal web parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Sweden. Uppsala University, Uppsala, Sweden, pp. 211–220.
- Mehdizadeh Seraj R., Siahbani M. and Sarkar A.** (2015). Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Association for Computational Linguistics, pp. 1379–1390.
- Pavlick E. and Kwiatkowski T.** (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics* 7, 677–694.
- Pivovarova L., Pronoza E., Yagunova E. and Pronoza A.** (2018). Paraphraser: Russian paraphrase corpus and shared task. In Filchenkov A., Pivovarova L. and Žižka J. (eds), *Artificial Intelligence and Natural Language*, Cham. Springer International Publishing, pp. 211–225.
- Reimers N. and Gurevych I.** (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 3982–3992.
- Scherrer Y.** (2020). TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 6868–6873.
- Schwenk H. and Douze M.** (2017). Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, Vancouver, Canada. Association for Computational Linguistics, pp. 157–167.
- Shimohata M., Sumita E. and Matsumoto Y.** (2004). Building a paraphrase corpus for speech translation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA), pp. 1407–1410.
- Soni S. and Roberts K.** (2019). A paraphrase generation system for EHR question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy. Association for Computational Linguistics, pp. 20–29.
- Tiedemann J.** (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA), pp. 2214–2218.
- Tiedemann J.** (2020). The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 1174–1182.
- Virtanen A., Kanerva J., Ilo R., Luoma J., Luotolahti J., Salakoski T., Ginter F. and Pyysalo S.** (2019). Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.

- Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O. and Bowman S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E. and Garnett R. (eds), *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.
- Wieting J. and Gimpel K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 451–462.

Cite this article: Kanerva J, Ginter F, Chang L-H, Rastas I, Skantsi V, Kilpeläinen J, Kupari H-M, Piirto A, Saarni J, Sevón M and Tarkka O (2024). Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for Finnish. *Natural Language Engineering* 30, 319–353. <https://doi.org/10.1017/S1351324923000086>