Open Subtitles Paraphrase Corpus for Six Languages

Mathias Creutz
Department of Digital Humanities, Faculty of Arts, University of Helsinki, Finland
mathias.creutz@helsinki.fi

Abstract

This paper accompanies the release of *Opusparcus*, a new paraphrase corpus for six European languages: German, English, Finnish, French, Russian, and Swedish. The corpus consists of paraphrases, that is, pairs of sentences in the same language that mean approximately the same thing. The paraphrases are extracted from the OpenSubtitles2016 corpus, which contains subtitles from movies and TV shows. The informal and colloquial genre that occurs in subtitles makes such data a very interesting language resource, for instance, from the perspective of computer assisted language learning. For each target language, the Opusparcus data have been partitioned into three types of data sets: training, development and test sets. The training sets are large, consisting of millions of sentence pairs, and have been compiled automatically, with the help of probabilistic ranking functions. The development and test sets consist of sentence pairs that have been checked manually; each set contains approximately 1000 sentence pairs that have been verified to be acceptable paraphrases by two annotators.

Keywords: paraphrase, subtitle, colloquial language, annotation, ranking, German, English, Finnish, French, Russian, Swedish

1. Introduction

This paper introduces the first release of the Opusparcus multilingual corpus of paraphrases (Creutz, 2018). Paraphrases are pairs of phrases in the same language that essentially convey the same meaning, such as *"Have a seat."* versus *"Sit down."*. Paraphrase resources have been published earlier, for instance by Quirk et al. (2004), Dolan et al. (2004), Dolan and Brockett (2005), Ganitkevitch et al. (2013), Ganitkevitch and Callison-Burch (2014), and Pavlick et al. (2015). However, Opusparcus has a few distinctive characteristics.

Firstly, and most importantly, all paraphrases in Opusparcus (**Op**en**S**u**b**titles**Par**aphrase**C**orp**us**) consist of movie and TV subtitles extracted from the *OpenSubtitles2016* collection of parallel corpora (Lison and Tiedemann, 2016). Previous paraphrase collections mostly contain fairly formal language in the form of news text and transcripts of parliamentary proceedings. The more *colloquial* language used in subtitles can be a valuable addition, for instance, in computer assisted language learning, to help learners find natural and idiomatic expressions in real-life situations.

Secondly, in this work the *pivot* language technique introduced by Bannard and Callison-Burch (2005) is applied using multiple pivot languages rather than just one or a few. The technique consists in finding paraphrases in one target language by translating to another, so-called pivot language and then translating back. For example, English *"Have a seat."* can be translated to French *"Asseyez-vous."*, which can be translated back to *"Sit down."*. Now, a well known fact is that different languages make different distinctions; for instance, the English pronoun *you* corresponds to French *toi* or *vous*, depending on number and degree of politeness. If French paraphrases are extracted using English as a pivot, then the *toi/vous* distinction will typically disappear, such that *"Asseyez-vous."* and *"Assieds-toi."* emerge as paraphrases, because they can both be translated as *"Sit down."*. Whether this is desirable or not depends on the application. However, if multiple pivot languages are used rather than one, more distinctions can be preserved. Bannard and Callison-Burch (2005) use four pivot languages in order to identify English paraphrases. Denkowski and Lavie (2010) use one, two, or three pivot languages for their five target languages. Ganitkevitch and Callison-Burch (2014) produce paraphrases for an impressive number of 21 languages, but they limit themselves to using one language, English, as their pivot (in order to be able to use syntactic information, which is available only for English). Opusparcus contains paraphrases in six European languages representing four different language branches: German, English, Swedish (Germanic), French (Romance), Russian (Slavic), and Finnish (Finnic). For each of the six languages, all other five languages are used as pivots.

Thirdly, simplicity is reflected in several aspects of the work. On one hand, only full sentences, so called sentential paraphrases, are produced, unlike Ganitkevitch et al. (2013), Ganitkevitch and Callison-Burch (2014), and Pavlick et al. (2015), who also extract sub-sentential paraphrases, such as individual word pairs, and include the counts of all such fragments in their reported figures. On the other hand, typically subtitles are fairly short, which makes it easier to evaluate and annotate the paraphrase candidates, unlike the complex sentences in the news data of Dolan and Brockett (2005). Furthermore, sub-sentential features or syntactic constraints (Callison-Burch, 2008) are not utilized to assess the likelihood that two sentences are paraphrases. If one favors similar sentence structures, there is a risk to miss some interesting idiomatic variation, such as in *"It's what we do."* $\leftrightarrow$ *"This is our job."*. Finally, particular to this work is that paraphrases and scores for ranking paraphrases are *symmetric*. The two phrases are equal, for instance in contrast to the incorporation of fine-grained entailment relations (Pavlick et al., 2015; Bowman et al., 2015) and the asymmetric conditional probabilities used by Bannard and Callison-Burch (2005).

The rest of this article is split into two main blocks, followed by some concluding remarks. The data sets and annotation scheme are described in Section 2. Alternative ranking functions that can be utilized to produce large paraphrase corpora are evaluated in Section 3.

Category	Description	Examples
Good *"Green"*	The two sentences can be used in the same situation and essentially "mean the same thing".	*It was a last minute thing.* ↔ *This wasn't planned.* *Honey, look.* ↔ *Um, honey, listen.* *I have goose flesh.* ↔ *The hair's standing up on my arms.*
Mostly good *"Light green"*	It is acceptable to think that the two sentences refer to the same thing, although one sentence might be more specific than the other one, or there are differences in style, such as polite form versus familiar form.	*Hang that up.* ↔ *Hang up the phone.* *Go to your bedroom.* ↔ *Just go to sleep.* *Next man, move it.* ↔ *Next, please.* *Calvin, now what?* ↔ *What are we doing?* *Good job.* ↔ *Right, good game, good game.*
Mostly bad *"Yellow"*	There is some connection between the sentences that explains why they occur together, but one would not really consider them to mean the same thing.	*Another one?* ↔ *Partner again?* *Did you ask him?* ↔ *Have you asked her?* *Hello, operator?* ↔ *Yes, operator, I'm trying to get to the police.*
Bad *"Red"*	There is no obvious connection. The sentences mean different things.	*She's over there.* ↔ *Take me to him.* *All the cons.* ↔ *Nice and comfy.*

Table 1: The four annotation categories used, with examples. Each category is also associated with a color, which corresponds to the color of a button in the user interface of the annotation tool.

2. Data Sets and Annotation Scheme

OpenSubtitles2016 (Lison and Tiedemann, 2016) is a collection of translated movie and TV subtitles from `www.opensubtitles.org`. OpenSubtitles2016, which is a subset of the larger OPUS collection[1], provides a large number of sentence-aligned parallel corpora in 65 languages. When subtitles exist for the same film in multiple languages, then sentence alignments are available for each language pair. For the present work, fifteen such bitexts were used, that is, all language-pair combinations for the six target languages German, English, Finnish, French, Russian and Swedish.

In principle, we work on full sentences only, and thus the terms *sentence* and *phrase* are used fairly interchangeably in this paper. Only one-to-one aligned sentences are used, that is, one sentence in the target language must be aligned with one sentence in the pivot language. There are occasional OCR errors and incorrect sentence segmentation in the data.

For each language, the data have been partitioned into separate training, development, and test sets, based on the release year of the movie; the test sets were extracted from years ending in 4, development sets from years ending in 5, and training sets from the rest.

Four different categories have been used when annotating sentence pairs. The annotation scheme is illustrated in Table 1. Other annotation schemes exist as well, such as slightly more complex, five-level Likert scales (Callison-Burch, 2008).

2.1. Training Sets

The so-called training sets are orders of magnitudes larger than the development and test sets and consist of lists of automatically ranked sentence pairs, where a high rank means a higher probability that the two sentences are paraphrases.

[1]OPUS ("... the open parallel corpus"): `opus.lingfil.uu.se`

The training sets, or subsets of them, are intended to be used freely for any useful purpose.

The training sets were produced as follows: First, around 1000 randomly selected sentence pairs were annotated by the author for each of the six languages. Then, an automatic ranking function was applied to all sentence pairs (annotated and unannotated alike), as explained later in Section 3. By extrapolating from the manually annotated data points to the entire set, an estimate of the quality of the training sets can be obtained. The result for English is shown in Figure 1.

Another view to the quality of the training sets is provided in Table 2, where approximate corpus sizes are given for each of the six languages, at three different accuracy levels.

Language	95%	90%	75%
German (de)	590,000	1,000,000	4,700,000
English (en)	1,000,000	1,500,000	7,000,000
Finnish (fi)	480,000	640,000	3,000,000
French (fr)	940,000	2,400,000	11,000,000
Russian (ru)	150,000	170,000	3,400,000
Swedish (sv)	240,000	600,000	1,400,000

Table 2: Number of phrase pairs in the training sets at three different cut-off points, where 95%, 90%, and 75% of the sentence pairs are estimated to be "Good" or "Mostly good" paraphrases.

2.2. Development and Test Sets

Whereas the training sets have been produced semi-automatically, the development and test sets consist exclusively of sentence pairs that have been annotated manually. This is to guarantee the high quality of these sets. However, quality comes at the expense of quantity, so the development and test sets are smaller than the training sets. The number of annotations produced for each language are shown i Table 3. Half of the sentence pairs belong to the

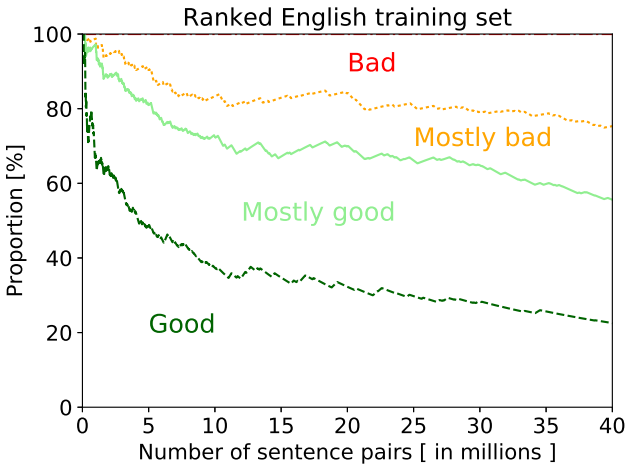

Figure 1: Estimated quality of the English training set. Proportions of the four annotation categories are calculated cumulatively, starting from the most highly ranked sentence pairs (on the left). By picking only highly ranked sentence pairs, one can achieve a high share of "Good" and "Mostly good" parapahrase candidates. The more data that is included, the more "Bad" or "Mostly bad" sentence pairs appear in the set.

development set and the other half to the test set.

Language	Total number of annotations	"Good" or "Mostly good" paraphrases
German (de)	3483	2060
English (en)	3088	1997
Finnish (fi)	3703	1921
French (fr)	3847	2004
Russian (ru)	4381	2088
Swedish (sv)	4178	1931
Total	22,680	12,001

Table 3: Total number of manual annotations in the development and test sets combined. Each sentence pair has been annotated independently by two annotators. For each language, approximately 2000 annotated sentence pairs qualify as acceptable paraphrases.

The development sets can be used to refine whatever training algorithms one might want to devise, and the test sets should be used in final evaluations only. The development sets contain only sentence pairs that do not occur in the training sets. The test sets consist of sentence pairs that do not occur in either the training sets or the development sets. The sentence pairs to be annotated manually were subject to more rigorous pre-filtering than the sentence pairs in the training sets. In the data, there are many sentences that differ only slightly from each other, such as: *"He is not your friend."* ↔ *"He isn't your friend."*. It would have been a waste of human labor to have such simple and predictable variations annotated manually. Therefore, only pairs of sentences that differ sufficiently from each other are accepted into the development and test sets. The difference is measured using relative edit distance; in general, the edit distance between the two sentences has to be at least 0.4 times the length of the shorter of the sentences (and for very short sentences containing less than 24 characters, the distance threshold is even higher).

Two persons annotated every sentence pair. If the annotators agreed on the category, the annotation was accepted as is. If the annotators disagreed but picked adjacent categories (such as "Good" versus "Mostly good" or "Mostly good" versus "Mostly bad"), then the annotation was also accepted, but the lower category was assigned (such that "Mostly good" and "Mostly bad" yields "Mostly bad"). If there was stronger disagreement between the annotators (such as "Mostly good" versus "Bad"), then the sentence pair was discarded. The annotators were also able to discard a sentence pair, if the language of either sentence was wrong or there were spelling or grammar errors. The number of trashed sentences turned out to be highest for French and Russian: It appears that French orthography is complex and mistakes are fairly common in written text. In the Russian data, some non-Russian Cyrillic as well as Latin characters show up occasionally, apparently because of inaccurate optical character recognition (OCR).

The detailed outcome of the annotation effort is summarized in Table 4 for the development sets and Table 5 for the test sets.

3. Automatic Ranking of Paraphrase Candidates

For the data sets that are intended to be used as training sets, a number of ranking schemes have been tested in order to identify paraphrases. Five of the ranking schemes are presented below, followed by a description how these approaches were evaluated. In the examples, English is used as our target language, and we are looking for English paraphrases. In the actual experiments, English was just one of the languages, and the same procedure was carried out for German, Finnish, French, Russian, and Swedish, as well.

3.1. Conditional Probability

Bannard and Callison-Burch (2005) propose a conditional paraphrase probability $P(e_2|e_1)$ as the probability that the English phrase e_1 is translated to a foreign phrase f_i, which in turn is translated back into another English phrase e_2. Since there are typically multiple possible foreign translations, we need to marginalize over the different possible f_i:

$$P(e_2|e_1) = \sum_i P(e_2|f_i)P(f_i|e_1) \quad (1)$$

This ranking formula tends to assign high ranks to phrase pairs, where e_1 is more specific than e_2. For instance, consider the case, where e_1 is *"I was taken from my family when I was a boy."* and e_2 is *"I was taken from my family."*. In the English-French parallel corpus, both English phrases have been aligned with the French phrase f_1: *"On m'a enlevé à ma famille."*. However, e_1 occurs aligned against f_1 only once, whereas e_2 21 times. Thus, $P(f_1|e_1)$ is high (=1), because e_1 is always translated as f_1. Also, $P(e_2|f_1)$ is high (= 21/22), because f_1 is almost always translated as e_2.

	Good	Mostly good		Mostly bad		Bad		Discarded	
	2 x Green	Gr. + Light gr.	2 x Light green	Light gr. + Yel.	2 x Yellow	Yellow + Red	2 x Red	Trash	Disagree
de	286 (16.4%)	333 (19.1%)	394 (22.6%)	189 (10.9%)	112 (6.4%)	100 (5.7%)	168 (9.6%)	81 (4.7%)	79 (4.5%)
en	409 (26.5%)	319 (20.7%)	287 (18.6%)	105 (6.8%)	74 (4.8%)	48 (3.1%)	213 (13.8%)	61 (4.0%)	28 (1.8%)
fi	351 (19.0%)	268 (14.5%)	344 (18.6%)	185 (10.0%)	135 (7.3%)	135 (7.3%)	342 (18.5%)	36 (1.9%)	56 (3.0%)
fr	252 (13.1%)	337 (17.5%)	408 (21.2%)	226 (11.7%)	207 (10.8%)	94 (4.9%)	106 (5.5%)	229 (11.9%)	65 (3.4%)
ru	473 (21.6%)	337 (15.4%)	210 (9.6%)	256 (11.7%)	155 (7.1%)	221 (10.1%)	202 (9.2%)	185 (8.4%)	152 (6.9%)
sv	376 (18.0%)	303 (14.5%)	305 (14.6%)	155 (7.4%)	86 (4.1%)	161 (7.7%)	501 (24.0%)	105 (5.0%)	97 (4.6%)

Table 4: Detailed breakdown of the results of the annotation of the development sets. A sentence pair qualifies as a "good" paraphrase, when both annotators have chosen the "good" category, visualized as a green button in the annotation tool. A sentence pair qualifies as "mostly good", when either one annotator has pushed the green button and the other annotator has pushed the light green button or both annotators have chosen the light green button. Similarly, sentence pairs have been categorized as "mostly bad" or "bad", if both annotators have agreed on the same category or if the annotators ended up pushing adjacent buttons. Sentence pairs were discarded in the following scenarios: The pair was trashed, if at least one of the annotators judged it to contain incorrect spelling or grammar. The sentence pair was also discarded, if the annotators disagreed about the category by more than one step on the four-level scale.

	Good	Mostly good		Mostly bad		Bad		Discarded	
	2 x Green	Gr. + Light gr.	2 x Light green	Light gr. + Yel.	2 x Yellow	Yellow + Red	2 x Red	Trash	Disagree
de	303 (17.4%)	333 (19.1%)	411 (23.6%)	177 (10.2%)	116 (6.7%)	85 (4.9%)	161 (9.2%)	77 (4.4%)	78 (4.5%)
en	450 (29.1%)	273 (17.7%)	259 (16.8%)	97 (6.3%)	56 (3.6%)	64 (4.1%)	246 (15.9%)	59 (3.8%)	40 (2.6%)
fi	376 (20.3%)	244 (13.2%)	338 (18.3%)	179 (9.7%)	138 (7.5%)	121 (6.5%)	353 (19.1%)	60 (3.2%)	42 (2.3%)
fr	261 (13.6%)	337 (17.5%)	409 (21.3%)	206 (10.7%)	204 (10.6%)	124 (6.4%)	133 (6.9%)	184 (9.6%)	65 (3.4%)
ru	462 (21.1%)	351 (16.0%)	255 (11.6%)	223 (10.2%)	151 (6.9%)	225 (10.3%)	188 (8.6%)	189 (8.6%)	146 (6.7%)
sv	379 (18.1%)	312 (14.9%)	256 (12.3%)	173 (8.3%)	107 (5.1%)	147 (7.0%)	527 (25.2%)	120 (5.7%)	68 (3.3%)

Table 5: Detailed breakdown of the results of the annotation of the test sets. Exactly the same procedure was applied as for the development sets. Annotators were unaware of which set a particular sentence pair belonged two; in fact, most annotators were unaware of the existence of separate development and test sets.

This tendency produces numerous errors, when there are occasional misaligned phrases in the corpus, such as in: *"We're staying in the army."* $\rightarrow$ *"Aah."*, where *"We're staying in the army."* has been aligned against French *"Aah."* once, which in turn has been aligned with English *"Aah."* 1401 times.

3.2. Joint Probability

Instead of a conditional probability, which is asymmetric, one can use the corresponding joint probability, which includes a prior probability, and is symmetric. Thus, the probability of e_1 being a paraphrase of e_2 is the same as the probability of e_2 being a paraphrase of e_1:

$$P(e_1, e_2) = P(e_2|e_1)P(e_1) = P(e_1|e_2)P(e_2) \quad (2)$$

$P(e_2|e_1)$ and $P(e_1|e_2)$ are calculated as in Equation (1), and $P(e_1)$ and $P(e_2)$ are the (prior) probabilities of the phrases, which are simply estimated as relative frequencies over all sentences in the corpus.

Now, at the top of the ranking, we find pairs consisting of frequently used phrases: *"Yes."* $\leftrightarrow$ *"Yeah."*, *"Of course."* $\leftrightarrow$ *"Sure."*, *"Hello."* $\leftrightarrow$ *"Good morning."*, *"Are you okay?"* $\leftrightarrow$ *"Are you all right?"*.

A few spurious phrase pairs also score high, where it appears that two frequent phrases might have found a common translation mostly by chance, by the fact that they occur frequently in general: *"You 're welcome."* $\leftrightarrow$ *"Sure."*, *"Yeah."* $\leftrightarrow$ *"I am."*, *"Hi."* $\leftrightarrow$ *"Goodbye."*, *"I do."* $\leftrightarrow$ *"I know."*

3.3. Pointwise Mutual Information

Pointwise Mutual Information (PMI) divides the joint probability by the probability that the two phrases $e_1$and e_2 occur independently. Thus, PMI penalizes phrase pairs that co-occur mostly by chance, by the fact that they occur frequently in general:

$$\begin{aligned} \text{pmi}(e_1; e_2) &= \log \frac{P(e_1, e_2)}{P(e_1)P(e_2)} \\ &= \log \frac{P(e_2|e_1)}{P(e_2)} = \log \frac{P(e_1|e_2)}{P(e_1)} \end{aligned} \quad (3)$$

This scoring favors phrase pairs e_1 and e_2 that have a limited set of translations f_i, such that e_1 and e_2 are not aligned with phrases other than f_i, and f_i are not aligned with other phrases than e_1 and e_2. For instance, the phrases *"You sound a little homesick."* $\leftrightarrow$ *"Do you miss being home?"* have a common French translation *"Vous avez le mal du pays ?"*, which occurs twice in the corpus, aligned once against each of the two English phrases.

However, similarly to the conditional probability in Equation (1), PMI is sensitive to misaligned, infrequent sentences. The phrase pair *"Lost the phone now."* $\leftrightarrow$ *"I'm from the agency."* scores high because *"Lost the phone now."*

has been misaligned against French *"Je viens de l'agence."*, which occurs only twice.

3.4. Joint Probability and PMI Combined

Our experiments show that rather than using joint probability (2) or PMI (3) in isolation, we obtain a better ranking by multiplying the two together:

$$P(e_1, e_2) \cdot \text{pmi}(e_1; e_2) = P(e_1, e_2) \log \frac{P(e_1, e_2)}{P(e_1)P(e_2)} \quad (4)$$

This leverages the strengths and alleviates the shortcomings of the two approaches.

3.5. Multiple Multilingual Parallel Corpora

The four formulae presented above, (1), (2), (3), and (4), easily generalize beyond bilingual parallel corpora. One can simply concatenate all parallel corpora, such that English is kept on one side and the other (pivot) languages on the other side. All frequencies and probabilities are then calculated over the merged bitext as a whole.

Interestingly, another approach can produce better results. PMI in Equation (3) may rank rare phrase pairs very high, and in case of misalignments, these pairs are unreliable. However, if a rare phrase pair ranks high in multiple bitexts, then this seems to signal much higher confidence. In order not to lose the information that a phrase pair emerges in multiple different corpora, rather than merging the parallel corpora into one, we can keep them separate. We then compute PMI scores separately for each bitext (English-German, English-Finnish, English-French, English-Russian, and English-Swedish). To obtain a combined score, we compute the sum of the PMI values obtained from each different corpus (English vs. pivot language $\mathcal{L}_i$):

$$\sum_i \text{pmi}(e_1; e_2|\mathcal{L}_i) = \sum_i \log \frac{P(e_1, e_2|\mathcal{L}_i)}{P(e_1|\mathcal{L}_i)P(e_2|\mathcal{L}_i)} \quad (5)$$

The probabilities are calculated exactly as previously. The notation merely highlights the fact that every value is conditioned on alignments between English and a specific pivot language $\mathcal{L}_i$.

Since the number of languages is constant, the sum in (5) can also be interpreted as the average PMI across pivot languages.

3.6. Evaluation of Ranking Schemes

A symmetric score is desired in this work, and therefore the conditional probability in Equation (1) cannot be used as such. However, one could obtain a symmetric score by combining $P(e_2|e_1)$ and $P(e_1|e_2)$ in some way, such as taking the minimum, maximum or average value. In practice, this would make this score behave fairly similarly as the more elegantly formulated PMI in Equation (3), so the conditional probability scheme was not investigated further. This leaves us with the four remaining schemes in Equations (2), (3), (4), and (5). They were compared with the help of a set of phrase pairs that were drawn randomly from the training set and annotated manually, as described in Section 2.1. The training set is then reordered using the ranking scheme to be tested. Depending on the ranking scheme, the manually annotated phrase pairs will appear at different ranks in the full, ordered collection. An ideal ranking scheme will place the phrase pairs that are true paraphrases at the head of the ordering and the phrase pairs that are not paraphrases at the tail.

The results were then plotted as in Figure 1 and compared visually. Across the six languages, the results were consistent: the best performing rankings were PMI summed over multiple corpora (5) followed by joint probability mutliplied by PMI (4). The types of phrase pairs that rank high are different in both cases: the former favors less frequent, more specific phrase pairs, such as *"It was a difficult and long delivery."* ↔ *"The delivery was difficult and long."*, whereas the latter favors frequent, less informative phrase pairs, such as: *"Excuse me."* ↔ *"I'm sorry."*. PMI summed over multiple corpora, in Equation (5), was judged to be the best ranking function. The final training sets were produced using this particular ranking.

4. Conclusion

Paraphrase extraction from movie subtitle data has been described in this paper. Six languages were included in this initial phase, but there is no principal reason why not more of the 65 languages in the OpenSubtitles2016 collection could be exploited. As there is considerable manual annotation effort involved, crowdsourcing could be considered; see, for instance, Tschirsich and Hintz (2013).

Another improvement could be to reduce the number of OCR errors that still occur in the data.

Acknowledgments

The following people have participated in the annotation effort: Thomas de Bluts, Aleksandr Semenov, Eetu Sjöblom, Mikko Aulamo, Olivia Engström, Janine Siewert, Carola Carpentier, Svante Creutz, Yves Scherrer, Anders Ahlbäck, Sami Itkonen, Riikka Raatikainen, Kaisla Kajava, Tiina Koho, Oksana Lehtonen, Sharid Loáiciga Sánchez, Tatiana Batanina, and the author himself. The annotation tool was implemented by Mikko Aulamo. The author is grateful to Mikko and the annotators for their very valuable contributions.

The author would also like to thank professor Jörg Tiedemann together with the anonymous reviewers for their valuable comments and suggestions as well as the Academy of Finland for the financial support of the annotations through Project 314062 in the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence.

Furthermore, warm thanks go to Mietta Lennes and the Language Bank of Finland (Kielipankki) for publishing and hosting Opusparcus as part of their corpus collections.

5. Bibliographical References

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 196–205, Honolulu, Hawaii. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2010). METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, WMT '10*, pages 339–342, 08.

Dolan, B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing, January.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Geneva, Switzerland. Association for Computational Linguistics.

Ganitkevitch, J. and Callison-Burch, C. (2014). The multilingual paraphrase database. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May. European Language Resources Association.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 425–430, Beijing, China, July. Association for Computational Linguistics.

Quirk, C., Brockett, C., and B. Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2004)*, pages 142–149, Barcelona, Spain, January.

Tschirsich, M. and Hintz, G. (2013). Leveraging crowdsourcing for paraphrase recognition. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse @ ACL 2013*, pages 205–213, Sofia, Bulgaria, August.

6. Language Resource References

Creutz, M. (2018). *Opusparcus: Open Subtitles Paraphrase Corpus for Six Languages,* version 1.0. Language Bank of Finland (Kielipankki), `http://urn.fi/urn:nbn:fi:lb-2018021221`.