Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6868–6873
Marseille, 11–16 May 2020

TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages

Yves Scherrer
University of Helsinki
Department of Digital Humanities, Faculty of Arts
yves.scherrer@helsinki.fi

Abstract

This paper presents TaPaCo, a freely available paraphrase corpus for 73 languages extracted from the Tatoeba database. Tatoeba is a crowdsourcing project mainly geared towards language learners. Its aim is to provide example sentences and translations for particular linguistic constructions and words. The paraphrase corpus is created by populating a graph with Tatoeba sentences and equivalence links between sentences "meaning the same thing". This graph is then traversed to extract sets of paraphrases. Several language-independent filters and pruning steps are applied to remove uninteresting sentences. A manual evaluation performed on three languages shows that between half and three quarters of inferred paraphrases are correct and that most remaining ones are either correct but trivial, or near-paraphrases that neutralize a morphological distinction. The corpus contains a total of 1.9 million sentences, with 200 – 250 000 sentences per language. It covers a range of languages for which, to our knowledge, no other paraphrase dataset exists. The dataset is available at `https://doi.org/10.5281/zenodo.3707949`.

Keywords: Multilingual corpus, Paraphrases, Crowdsourcing

1. Introduction

Paraphrases are different textual realizations of the same meaning within a single language (Bhagat and Hovy, 2013; Ganitkevitch and Callison-Burch, 2014). Paraphrase detection and generation have become popular tasks in NLP and are increasingly integrated into a wide variety of common downstream tasks such as machine translation, information retrieval, question answering, and semantic parsing (Federmann et al., 2019). Although the availability of large datasets for training and evaluation has facilitated research on paraphrase detection and generation, most of these datasets cover only a single language – in most cases English – or a small number of languages (see Section 4.). Furthermore, some paraphrase datasets focus on lexical and phrasal rather than sentential paraphrases, while others are created (semi-)automatically using machine translation.

This paper describes the creation of a paraphrase corpus for 97 languages with a total of 1.9 million sentences. It consists of entire sentences produced by crowdsourcing within the Tatoeba project (Section 2.). The paraphrase matching process is entirely automatic and is based on the multilingual pivoting approach introduced by Bannard and Callison-Burch (2005). The number of sentences per language ranges from 200 to 250 000, which makes the dataset more suitable for fine-tuning and evaluation purposes than for training. It covers languages for which, to our knowledge, no other paraphrase dataset exists.

In contrast to some previous work, we organize paraphrases as sets rather than pairs: all sentences in a *paraphrase set* are considered paraphrases of each other. This representation is especially well-suited for multi-reference evaluation of paraphrase generation models, as there is generally not a single correct way of paraphrasing a given input sentence.

This paper is structured as follows: Section 2. describes the Tatoeba project, from which the data is taken. Section 3. describes the different steps involved in the creation of the multilingual paraphrase corpus. Section 4. compares TaPaCo with related resources and discusses some limitations of the TaPaCo creation process. Section 5. reports on the results of a manual evaluation of a small selection of the corpus.

2. The Tatoeba project

Tatoeba identifies itself as "a large database of sentences and translations".[1] The Tatoeba database is populated by crowdsourcing: anybody can propose sentences in any language, and anybody can propose translations of existing sentences into another language. As a result, several translations of the same sentence into the same language are common, and we exploit these alternative translations as paraphrases. Furthermore, contributors can add sentences to lists[2] and annotate them with tags.[3]

The Tatoeba project was started in 2006 with the intention to provide example sentences for particular linguistic constructions and words to help language learners.[4] Hence, the material provided by Tatoeba consists mainly of simple, short sentences in colloquial style. All data published on Tatoeba is released under CC-BY 2.0 FR.[5]

As a starting point, we use the Tatoeba dataset made available as part of the OPUS collection (Tiedemann, 2012).[6] OPUS provides sentence alignments for all available language pairs. We do not currently use other annotations provided by OPUS (tokenization, word alignment, parsing).

The OPUS version of Tatoeba covers 338 languages and contains a total of 7.8 million sentences. Sentence alignment information is available for 1679 language pairs.

[1] `https://tatoeba.org/eng/about`

[2] For example, to specify the original source of the data; cf. `https://tatoeba.org/eng/sentences_lists/index`.

[3] For example, to indicate morphological or phonological properties of the sentence; cf. `https://tatoeba.org/eng/tags/view_all`.

[4] Tatoeba is Japanese and means 'for example'.

[5] `https://tatoeba.org`

[6] `http://opus.nlpl.eu/`

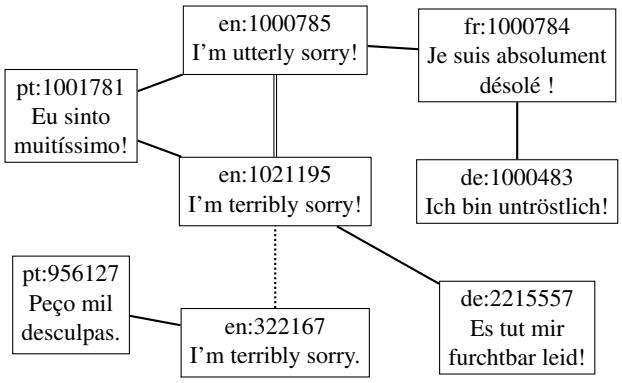

Figure 1: Small extract of the merged graph. Intra-language links are shown with double lines, surface similarity links with dotted lines and inter-language links with plain lines. The figure illustrates (a) that the two Portuguese sentences can only be connected thanks to the English sentence similarity link, and (b) that the two German sentences can only be connected by pivoting through multiple languages (English and French).

Moreover, OPUS provides intra-language sentence alignments for 28 languages.

3. Creating a multilingual paraphrase corpus

The process of building the TaPaCo dataset consists of three consecutive steps: first, we build a graph in which sentences of the same meaning are linked together, then traverse it to create sets of paraphrases, and finally remove uninteresting sentences from the sets.

3.1. Create equivalence graphs

In a first step, we create three graphs in which we introduce Tatoeba sentences as vertices[7] and links between equivalent sentences as (undirected) arcs. Each graph considers a different type of equivalence:

1. **Intra-lingual alignment links** connect sentences of the same language that have been identified as paraphrases by the OPUS sentence alignment process.
2. **Inter-lingual alignment links** connect sentences of different languages that have been identified as translations of each other by OPUS.
3. **Surface similarity links** connect sentences of the same language that become identical after punctuation normalization. Normalization maps typographic punctuation signs to plain ones, removes quotation marks and replaces exclamation marks by full stops.

These three graphs are then merged. Figure 1 provides an example of the merged graph with different types of edges. Table 1 shows the statistics of the graphs.
The intra-lingual links provide exactly the type of equivalence we are looking for, but they only cover a small number of languages and sentences. For this reason, we add inter-lingual links, which are at the core of the pivoting idea (Bannard and Callison-Burch, 2005). For example, inter-lingual links allow us to connect two English sentences through a common Italian translation. While previous work has explored only small numbers of pivoting languages, we use all languages available in Tatoeba for pivoting and also allow paths that go through more than one pivoting node. For example, two German sentences may be connected by a path through English and French sentences (see Figure 1). Finally, the surface similarity links are not particularly interesting by themselves (and will be removed again later), they are merely introduced to increase the connectivity of the merged graph.

[7] In rare cases, vertices may consist of more than one sentence. For simplicity, we continue referring to a vertex as "sentence".

Equivalence type	Languages	Vertices	Arcs
1 (intra-lingual links)	28	272k	159k
2 (inter-lingual links)	309	6 776k	7 724k
1 ∪ 2	309	6 891k	7 882k
3 (surface similarity)	137	45k	23k
1 ∪ 2 ∪ 3	310	6 893k	7 903k

Table 1: Statistics of the equivalence graphs. Each vertex contains a sentence, and each arc represents a link between two sentences.

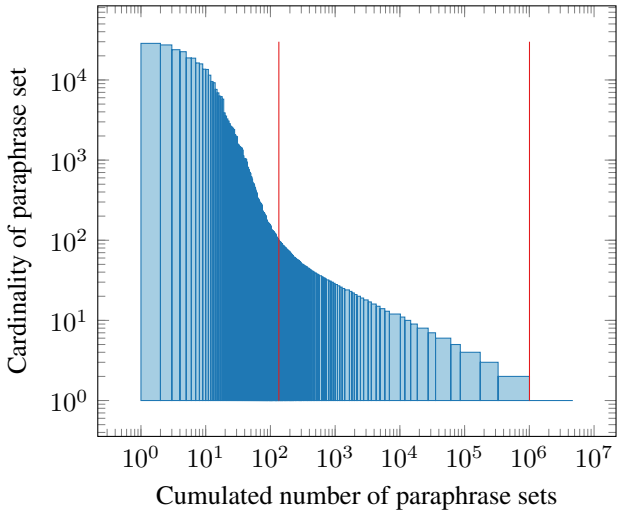

Figure 2: Distribution of paraphrase sets according to their cardinality. The pruning thresholds described in Section 3.2. are visualized by the vertical red lines.

3.2. Create paraphrase sets

The merged graph covers almost 7 million sentences in more than 300 languages. This graph is not fully connected. Rather, it consists of almost 1 million connected subgraphs, each of which represents a set of equivalent sentences in multiple languages. These subgraphs are then broken up by language such that all sentences of the same language end up in the same *paraphrase set*.
The cardinalities of the paraphrase sets follow a Zipfian distribution (see Figure 2): while the largest sets contain several thousand sentences, a large number of sets only contain a single sentence. We have chosen to discard both extremes:

Filter	Languages	Paraphrase sets	Sentences
Initial	310	4 698 620	6 893 427
Singleton sets	299	1 005 034	3 199 841
>100 sentences	299	1 004 899	2 834 100
Near-identical	298	995 979	2 796 128
BLEU filter	297	726 186	1 936 848
<100 sets/lang.	73	723 486	1 927 077

Table 2: Statistics of paraphrase sets. The horizontal line separates the filters described in Section 3.2. from those of Section 3.3.

- Singleton sets, i.e. sets containing only one sentence, are useless for paraphrase-related tasks and are therefore removed.
- Paraphrase sets with more than 100 sentences are likely to be of low quality, either because the meaning of the sentences is underspecified, or because of alignment errors that erroneously connect subgraphs of different meanings. These paraphrase sets are also removed.

The effects of these pruning steps can be seen in Table 2 (upper half).

3.3. Remove near-identical sentences

Initial checks have shown that a large number of paraphrase sets contain sentences that differ only in tokenization, punctuation or casing. We discard such "uninteresting" cases in the following way. All sentences of a paraphrase set are normalized by lowercasing, applying Unicode character normalization and removing all punctuation and spacing. If two or more normalized sentences match, only the one with the lowest ID number is retained.[8]

This still leaves a large amount of paraphrase pairs with very similar syntactic structure, for example pairs contrasting American and British spelling, or contracted and non-contracted verb forms. To discard such uninteresting pairs, we therefore compute sentence-level BLEU scores (Papineni et al., 2002) between all pairs and remove the second sentence whenever the BLEU score is higher than 50.

We estimate that a meaningful evaluation of a paraphrase detection or generation system requires at least 100 paraphrase sets, i.e. at least 200 potential source sentences. Therefore, languages whose paraphrase set coverage lies below this threshold are removed from the dataset as well. Table 2 (lower half) shows the effects of these additional filters. Only 24% of the initial languages, 15% of paraphrase sets and 28% of sentences satisfy all requirements. Detailed statistics per language are listed in Table 3.

3.4. Format

The paraphrase dataset is made available under the same CC-BY 2.0 licence as the original material.[9] The dataset consists of one tab-separated file per language containing one sentence per row. Sentences with the same paraphrase set id are considered paraphrases.[10] Figure 3 shows an excerpt of the English file.

Lang.	PS.	Sent.	Lang.	PS.	Sent.
af	139	307	kab	5552	15944
ar	2708	6446	ko	224	503
az	284	624	kw	544	1328
be	646	1512	la	2814	6889
ber	22009	67484	lfn	1041	2313
bg	2639	6324	lt	3466	8042
bn	516	1440	mk	6080	14678
br	723	2536	mr	6956	16413
ca	235	518	nb	489	1094
cbk	118	262	nds	1107	2633
cmn	5100	12549	nl	9441	23561
cs	2832	6659	orv	192	471
da	4770	11220	ota	199	486
de	**48308**	**125091**	pes	1875	4285
el	3812	10072	pl	9109	22391
en	**62045**	**158054**	pt	29949	78430
eo	**78405**	**207848**	rn	290	648
es	32691	85064	ro	941	2092
et	113	241	ru	**91240**	**251263**
eu	257	573	sl	310	706
fi	12082	31753	sr	3237	8175
fr	**44195**	**116733**	sv	2941	7005
gl	167	351	tk	528	1165
gos	122	279	tl	458	1017
he	25201	68350	tlh	1195	2804
hi	848	1913	toki	1693	3738
hr	231	505	tr	**56338**	**142229**
hu	26066	67964	tt	1107	2398
hy	240	603	ug	470	1183
ia	1111	2548	uk	21209	54431
id	667	1602	ur	117	252
ie	218	488	vi	428	962
io	207	480	vo	143	328
is	724	1641	war	145	327
it	**62881**	**198919**	wuu	188	408
ja	16764	44267	yue	247	561
jbo	1149	2704			

Table 3: Paraphrase sets (PS) and sentences per language. Languages are abbreviated using ISO 639-1 and 639-2 codes. Subcorpora with more than 40 000 paraphrase sets or 100 000 sentences are highlighted in bold.

4. Related resources

A large number of paraphrase corpora have been proposed over the last years. In this section, we enumerate the most relevant corpora and compare them with TaPaCo.

MSRPC The Microsoft Research Paraphrase Corpus (Dolan et al., 2004; Dolan and Brockett, 2005) consists of

[8] Note that this normalization step removes, among others, all near-duplicates introduced by the "surface similarity links" described in Section 3.1..

[9] The dataset is available at `https://doi.org/10.5281/zenodo.3707949`.

[10] The paraphrase set ids are kept constant across languages. For example, the English sentences of set 40 are translations of the French sentences of set 40.

1313	297738	He isn't my cousin.	907;4000;7360;7417	SVC; 6 syllables; present simple
1313	574175	He's not my cousin.	907;4000;7361;7417	5 syllables
1892	2218079	You're among friends.	907;4000;7361;7389; 7412;8511;9026	4 syllables; present simple
1892	2218509	You're with friends.	907;4000;7389;7409; 9053	3 syllables; present simple
1892	5013251	You guys are among friends.	6905	
24158	1481280	Cat got your tongue?	649	ellipsis
24158	3174733	Did the cat get your tongue?		
24158	3174734	Has the cat got your tongue?		
842729	906575	It is nevertheless a good sentence.	921;4481	
842729	906578	It's a good sentence, anyway.	921	
842770	906758	What's your favorite flavor of ice cream?	907;4000	
842770	906785	What's your favorite ice cream flavor?	907;4000	9 syllables

Figure 3: Excerpt of the English paraphrase file. The first column shows the **paraphrase set id**, a running number that groups together all sentences that are considered paraphrases of each other. The second column contains the OPUS **sentence id**. The two rightmost columns contain sentence-level metadata from Tatoeba: **lists** and **tags**, respectively.

5801 pairs of English sentences. The sentences are automatically extracted from news corpora and rated for paraphrase quality by humans.

PPDB The Paraphrase Database[11] is an ongoing effort to provide paraphrases automatically extracted from parallel corpora. The PPDB contains lexical and phrasal paraphrases instead of sentential ones. The first version (Ganitkevitch et al., 2013) covers English and Spanish. An expanded version is available for 23 languages (Ganitkevitch and Callison-Burch, 2014). The second version (Pavlick et al., 2015) provides improved rankings and adds a range of annotations, but these improvements are only available for the English version. All databases are available in different sizes, with the smallest size only containing the most reliable paraphrase pairs.

QQP The Quora Question Pairs corpus (Iyer et al., 2017) contains pairs of equivalent questions extracted from the Quora website. The corpus covers a variety topics but is limited to questions in English.

Opusparcus The Open Subtitles Paraphrase Corpus (Creutz, 2018) contains ranked sentential paraphrase pairs extracted from movie subtitles. Its content is thus less formal and more colloquial in style. The corpus is available for six languages, which partially overlap with the largest subcorpora of TaPaCo.

ParaNMT-50M The particularity of this corpus (Wieting and Gimpel, 2018) is its construction: instead of matching existing sentences, this corpus pairs English sentences written by humans with English sentences produced by machine translation from Czech. It covers a wide variety of textual domains and is also one of the biggest datasets (50M sentences), thanks to its automatic creation.

Multilingual whispers (Federmann et al., 2019) focuses on informal English data (casual online conversations and e-mails). Its creation procedure is most similar to ParaNMT-50M in that paraphrases are explicitly created using a range of techniques (manual translation or paraphrasing by crowdsourcing workers or experts, machine translation). In total, 14 500 sentences are collected and annotated with their creation technique and with quality scores.

The main advantages of TaPaCo compared to the existing resources are its multilingual coverage (more than 4 times as many languages as the PPDB), and its focus on complete sentences written by humans. TaPaCo also has its limitations, which are discussed in more detail below.

4.1. Limitations of TaPaCo

Data source The sentences collected in Tatoeba tend to be short and structurally simple. Most sentences are constructed examples that rarely occur in everyday language. It is also possible that some sentences are created by non-native speakers of the language. Whether these aspects constitute a force or a weakness depends mainly on the application scenario. Due to the wide range of possible applications, we also refrain from providing fixed training/test splits.

No restrictions on paths The graph traversal procedure groups together all sentences of the same language that are connected by a path, without taking into account the length of the path, the nature of arcs on the path, and the number of pivot languages on the path. It may be argued that these properties of the path influence the quality of the obtained paraphrases. We currently do not have reliable evidence of such influence, which is why we refrain from implementing any restrictions on the accepted paths.

No paraphrase quality estimation Many existing datasets provide some quality score for each paraphrase pair. This score is either computed from heuristics or obtained from a classifier trained on manually annotated material. Some datasets even identify recurrent paraphrasing patterns using syntactic analysis. We are currently unable to provide such measures for the whole range of languages covered by the dataset.

Morphological neutralization As reported by Ganitkevitch and Callison-Burch (2014), the pivoting approach for finding paraphrases is prone to neutralization of morpho-

[11] http://paraphrase.org/

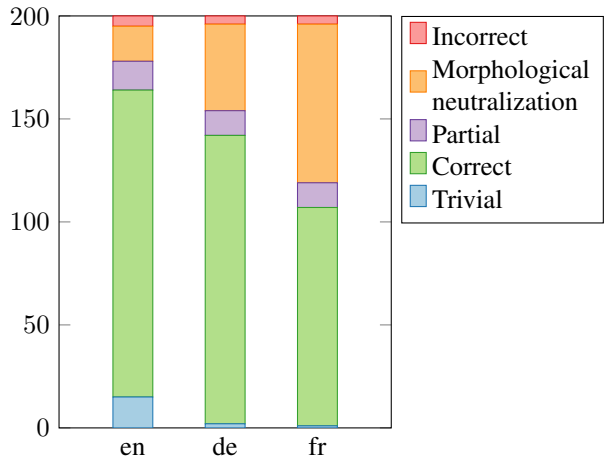

Figure 4: Results of the manual evaluation of 200 paraphrase pairs per language.

logical features that are not encoded explicitly in the pivot language. For example, the German sentences *Geh!* and *Gehen Sie!* encode different levels of politeness, but they are translated by the same English sentence *Go!* since English does not distinguish politeness levels. As a result, the two German sentences will be assumed to be paraphrases. Likewise, the English sentences *He is tired.* and *She is tired.* end up as paraphrases if they are linked by a pivot language that does not distinguish gendered pronouns (such as Finnish or Turkish). This phenomenon also appears in TaPaCo and could be alleviated by applying additional filters, potentially based on vector similarity between sentences. However, as Ganitkevitch and Callison-Burch (2014) state, "it is unclear whether this grouping is desirable or not, and the answer may depend on the downstream task."

5. Evaluation

In order to estimate the quality of the TaPaCo resource, we manually evaluated 200 paraphrase pairs from three languages with high coverage, namely English, German and French. We first randomly selected 200 paraphrase sets, and then randomly chose two sentences from each set.
Each paraphrase pair was annotated by the author with one of the following labels:

Correct but trivial Some paraphrase pairs are very similar and relate to each other by a regular pattern.

In English, the most frequent patterns were contracted vs. non-contracted verb forms (cf. set 1313 in Figure 3), and presence vs. absence of the *that* complementizer. In German, the patterns were related to spelling reform and to presence or absence of the *dass* complementizer. In French, the alternation between *on* and *nous* was considered as such a pattern.

Correct for sentence pairs that can be considered correct paraphrases of each other (e.g. sets 842729 and 842770 in Figure 3).

Partial for sentence pairs that have an evident semantic relation but which cannot be strictly considered paraphrases (e.g. *The food was great in Italy.* vs. *The Italian food was delicious.*).

Morphological neutralization This label is assigned to sentences which would be paraphrases of each other if one or more morphological features were considered to be irrelevant.

Incorrect for sentence pairs that cannot be considered to be paraphrases of each other (e.g. *Does Tom have a plan?* vs. *Does Tom have a piano?*[12]).

Figure 4 shows the results of the manual evaluation. It can be seen that English shows a larger proportion of trivial paraphrases, whereas German and French contain more cases of morphological neutralization. This mainly concerns the politeness feature, which is present in German and French but absent in English. In all three languages, more than half of all paraphrase pairs were evaluated as correct. Partial and incorrect paraphrases account for less than 10 percent of evaluated pairs.

6. Conclusion

We have presented TaPaCo, a freely available multilingual paraphrase corpus extracted from the Tatoeba database. It covers 73 languages – more than any other paraphrase dataset – with at least 100 paraphrase sets per language. The grouping of sentences into paraphrase sets (rather than paraphrase *pairs*) makes the dataset well-suited for multi-reference evaluation setups.
A manual evaluation of three high-coverage languages suggests that between half and three quarters of the paraphrase sets are correct. The majority of the remaining examples are either trivially correct or show some case of morphological neutralization. Future work will focus on marking these examples automatically using language-specific patterns or rules.

7. Acknowledgements

This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

8. Bibliographical References

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

Bhagat, R. and Hovy, E. (2013). Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[12] It is likely that this paraphrase pair was inferred by pivoting through the polysemous Italian word *piano*.

9. Language Resource References

Creutz, M. (2018). Open Subtitles paraphrase corpus for six languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Federmann, C., Elachqar, 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)*, pages 17–26, Hong Kong, China, November. 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 the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Iyer, S., Dandekar, N., and Csernai, K. (2017). First Quora dataset release: Question pairs.

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 (Volume 2: Short Papers)*, pages 425–430, Beijing, China, July. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.