A Multilingual Dataset for Evaluating Parallel Sentence Extraction from Comparable Corpora

Pierre Zweigenbaum
LIMSI, CNRS,
Université Paris-Saclay,
F-91405 Orsay, France
pz@limsi.fr

Serge Sharoff
University of Leeds
Leeds, United Kingdom
s.sharoff@leeds.ac.uk

Reinhard Rapp
Magdeburg-Stendal University
of Applied Sciences and
University of Mainz, Germany
reinhardrapp@gmx.de

Abstract

Comparable corpora can be seen as a reservoir for parallel sentences and phrases to overcome limitations in variety and quantity encountered in existing parallel corpora. This has motivated the design of methods to extract parallel sentences from comparable corporad. Despite this interest and work, no shared dataset has been made available for this task until the 2017 BUCC Shared Task. We present the challenges faced to build such a dataset and the solutions adopted to design and create the 2017 BUCC Shared Task dataset, emphasizing issues we had to cope with to include Chinese as one of the languages. The resulting corpus contains a total of about 3.5 million distinct sentences in English, French, German, Russian, and Chinese, mostly from Wikipedia. We illustrate the use of this dataset in the shared task and summarize the main results obtained by its participants. We finally outline remaining issues.

Keywords: Comparable corpora, parallel sentences, parallel sentence extraction, cross-language similarity, annotated corpus

1. Parallel Sentence Extraction from Comparable Corpora

Parallel sentences are the fuel needed to train machine translation systems. Large parallel corpora have been obtained from international bodies or collected from the Web. However, they only cover a small subset of the variety of language pairs, domains and genres that are found in language. Besides, because by construction at least half of the sentences in these corpora are the result of (human) translation, they are likely to display translation biases such as calques and other such phenomena.

Comparable corpora are (typically multilingual) corpora selected with similar criteria such as domain, genre, time period. In contrast to parallel corpora, they display much more variety and are normally original texts rather than translations. They hold much promise therefore as a complement to parallel texts for machine translation and other applications.

One way in which comparable corpora have been used to help machine translation is by spotting parallel sentences that occur naturally in these corpora, and using these sentence pairs to extend parallel corpora (Munteanu et al., 2004). This has motivated research into methods that aim to perform this task, such as (Utiyama and Isahara, 2003; Munteanu et al., 2004; Abdul-Rauf and Schwenk, 2009; Smith et al., 2010). This task is usually called *Parallel Sentence Extraction from Comparable Corpora*.

It is however difficult to compare earlier work and assess progress because of the absence of a shared dataset with gold standard annotations. Some past shared tasks addressed related objectives. Cross-language plagiarism detection in PAN (Potthast et al., 2012) aims to spot text that has been translated into a target language and reused in (inserted into) text in that target language. It is therefore quite close to our task. However, plagiarism detection can take advantage of differences in style between the original target text and the translated text, and of intrinsic properties of 'translationese'. This is not the case in our task, where all sentences are expected to be original. Cross-language text similarity as in SemEval 2016 (Agirre et al., 2016) assesses the level of semantic similarity of pairs of sentences on a given scale. It is also close to our task. Nevertheless, it has been proposed with already paired sentences instead of large monolingual corpora, thus removing the sentence spotting stage. Bilingual document alignment in a large Web collection has been proposed in WMT 2016 (Buck and Koehn, 2016). However, on the one hand it addressed documents instead of sentences; and on the other hand, it included meta-information in the form of document URLs, a property that we want to avoid.

This highlights the need for a publicly available dataset that would make it possible to compare methods that extract parallel sentences from comparable corpora. This paper describes the principles according to which we designed such a corpus, their implementation, the resulting corpus and a first use of that corpus in a shared task. This corpus was built in the context of the BUCC 2017 Shared Task described in (Zweigenbaum et al., 2017). The present paper provides more detail about our motivation and design criteria, about the rationale we followed to implement these design criteria, and about the processing of the Chinese part of the corpus.

2. A Dataset for Parallel Sentence Extraction from Comparable Corpora

2.1. Desiderata for a Dataset for the Task

We aimed to build a bilingual corpus to measure progress on the identification of parallel sentences in monolingual corpora. This led us to the following desiderata and design choices.

No metadata. We wish to focus on the cross-language comparison of sentence contents. Instead, most work so far has relied to a more or less large extent on meta-information: belonging to paired documents such as

linked Wikipedia pages or Web pages, sharing images or links to external documents, news published in a close time frame, etc. Our target corpus should not give such clues, and should therefore include no metadata on the documents in which the sentences are found.

Realistic size. Spotting parallel sentences is useful when performed at scale. We therefore aimed at a corpus with at least millions of sentences: although not especially large by today's standards, this already requires to use scalable algorithms.

Natural text. We wished to provide natural text rather than a simple list of sentences. First, this makes the task more realistic, since extraction of sentences from comparable corpora happens in the context of complete documents. Second, we can expect that the document context is likely to influence assessment of comparability between the sentences. This desideratum is probably less necessary, but we considered it would make the corpus closer to what the task should address.

Known true positives. Since we want to be able to evaluate system results, we must have a gold standard. This is the most challenging part of building such a corpus: recall that ultimately we want systems to spot pairs of sentences that occur naturally in a pair of monolingual corpora and happen to be translations of each other. We know of no such situation in which such sentence pairs would be marked in some way.

2.2. Pragmatic Choices

2.2.1. Creating an Artificial Corpus

Spotting naturally occurring sentence pairs in comparable corpora, if performed by humans, can be extremely time consuming: exhaustively spotting such pairs in, say, two corpora of 400,000 sentences each may require the examination of 160 billion sentence pairs. When preparing a gold standard is not feasible a priori, some shared tasks, e.g. ad hoc information retrieval, have resorted to pooling of system results then a posteriori human evaluation. A posteriori, we know that BUCC 2017 Shared Task participants produced a few hundred thousand sentence pairs: this is many orders of magnitude below that of the above-mentioned a priori evaluation, but is still sizeable. We did not have human resources to allocate to such a human evaluation either. We therefore decided to design a synthetic corpus containing controlled parallel sentences. We performed this by inserting known parallel sentence pairs into existing monolingual corpora.

We chose Wikipedia articles (20161201 dumps [1]) as our monolingual corpora and News Commentary (v11[2]) as our source for parallel sentence pairs.

In terms of domains, although in principle Wikipedia covers all domains, it over-represents named entities, specifically contemporary people and locations. The domain of News Commentary is that of commentaries on international news, hence it mentions a large number of contemporary people and locations. This results in a reasonable match of the domains of the two corpora.

Wiki genres	**%**	**NC genres**	**%**
Encyclopedic	66.4%	Argumentative	84.3%
Hard news	16.9%	Academic	2.6%
Argumentative	4.2%	Hard news	1.9%
Reviews	2.7%	Personal	1.6%
Academic	2.5%	Encyclopedic	1.4%

Table 1: Most common genres in our sources

We also performed a quantitative analysis of the two datasets in terms of topics and genres. Extraction of keywords and comparison of the cosine similarity between the resulting vectors (Sharoff, 2013) gives an estimate of how similar the documents are across the corpora. The inter-decile range of the cosine similarity scores between the News Commentary texts and their nearest Wikipedia counterparts is $[0.971, 0.980]$, i.e., for any News Commentary text it is nearly guaranteed that there is a sufficiently similar Wikipedia text.

Using the genre classifier from (Sharoff, 2018) we also assessed the genre composition of the two corpora, see Table 1. Even though precision of automatic genre classification for different genres varies from 65% to 85%, the results indicate a general trend confirming that the News Commentary corpus corresponds of a considerable portion of Wikipedia in terms of genres, albeit with a different distribution. Wikipedia contains many news-like or argumentative texts, which are similar to the News Commentary corpus, while the latter also contains some encyclopedic introductions and research-like texts similar to those found in Wikipedia.

The following two examples illustrate the similarities in both topics and genres between the two sources: Wikipedia id=13811803 *"Saltwater Keynesian economists" argue that business cycles represent market failures, and should be counteracted through discretionary changes in aggregate public spending and the short-term nominal interest rate. "Freshwater economists" often reject the effectiveness of discretionary changes in aggregate public spending as a means to efficiently stabilize business cycles.*

News Commentary: *The Chicago School claims that real-world market economies produce roughly efficient (so-called "Pareto optimal") outcomes on which public policy cannot improve. Thus, any state intervention in the economy must make someone worse off. The MIT School, by contrast, argues that real-world economies are afflicted by pervasive market failures, including imperfect competition and monopoly, externalities associated with problems like pollution, and an inability to supply public goods such as street lighting or national defense.*

In the remainder of this section we use French and English as a running example of a language pair. For convenience, we often call 'monolingual sentence' a sentence found in the monolingual corpora and 'parallel sentence' a sentence from the parallel corpora.

[1] http://ftp.acc.umu.se/mirror/wikimedia.org/dumps/

[2] http://www.casmacat.eu/corpus/news-commentary.html

2.2.2. Inserting Parallel Sentences in Monolingual Corpora

The inserted parallel sentence pairs should not be trivially detectable in the monolingual corpora. In other words, these sentences should be coherent with the context in which they are inserted. We aimed at topical coherence by looking in the monolingual corpora for sentences with similar contents to the parallel sentences and using the spotted sentences as insertion points. To perform this efficiently, we used a search engine to index each English sentence of the monolingual corpus (English Wikipedia dump, converted to text and split into sentences) and each French sentence of the monolingual corpus (French Wikipedia dump, converted to text and split into sentences). We used the Solr search engine and queried it for each sentence pair in the parallel corpus (French-English News Commentary) to find the most similar French sentence and English sentence for this pair.[3] If two similar enough sentences were found, we inserted the English parallel sentence after the matching English sentence in the monolingual corpus and the French parallel sentence after its matching French sentence in the monolingual corpus. Similarity constraints were enforced on the one hand by the Solr query parameters and on the other hand by post-filters including: a length in words (before stopword removal) in the range $[10, 20]$; a length ratio in the range $[0.8, 1.2]$.

In early experiments, we observed that the cohesion of the resulting sequence of two sentences was better if the start of the inserted sentence contained the common words with the pre-existing monolingual sentence (those words that make the sentences similar). To favor this, we decided to truncate the queries built from parallel sentences to the first T words. T was set to 5 words based on observations in these experiments.

2.2.3. Making Inserted Sentences Less Conspicuous

In early experiments, we realized that very short parallel sentences might happen to be inserted among much larger monolingual sentences, or the reverse. This would increase the risk that such sentences might break the cohesion of the original text. To reduce this risk, we acted on the distribution of sentence lengths in both the monolingual and parallel sentences: we excluded sentences shorter than 10 words and longer than 20 words. This range of lengths covered a large percentile of the original sentences, and is typical of what Machine Translation systems address.

We also realized that due to construction idiosyncrasies, Wikipedia texts had specific distributions of typographical features such as the presence of some typographical quotation marks; they were also subject to conversion issues that created systematic clues of their origin. This was notably caused by the use of Wikipedia templates, that were particularly numerous in the French Wikipedia. We endeavored to remove such idiosyncrasies by revisiting the Wikipedia conversion workflow. We added our own extensions to an existing Wikipedia conversion tool, WikiExtractor.py,[4]. We included sentence splitting based on NLTK, and removed the sentences that contained a Wikipedia template.

Removing sentences brought the additional advantage of making the original text slightly less cohesive: in that context of slightly reduced cohesion, the potential cohesion issues incurred by the addition of (parallel) sentences were likely to be less noticeable.

2.2.4. Controlling Unknown True Positives

The insertion of known parallel sentences aimed to control the true positives present in the datasets we were building. Parallel sentences might however already exist in the pair of monolingual corpora we started from. Indeed, the true nature of the task would be to find these pre-existing, naturally occurring parallel sentences. But we were instead aiming to populate our monolingual corpora with known parallel sentence pairs. We therefore needed to prevent as much as possible naturally occurring parallel sentence pairs from remaining in our monolingual corpora. The strategy we adopted in this purpose was to desynchronize our comparable corpora. Since we started from Wikipedia articles in two languages, we knew that interlinked articles would be highly likely to contain such parallel sentences: this is indeed a property that is often desired by past work on parallel sentence extraction. This is also how our previous shared task on detection of comparable texts has been setup (Sharoff et al., 2015): the gold standard was based on the iwiki links.

In contrast to such work, we built pairs of monolingual corpora which never contained two interlinked Wikipedia articles. This was also in line with our desideratum not to include meta-information on the sentences, such as being found in two interlinked articles.

The main drawback in doing so is that the most comparable pairs of documents for a given language pair are removed from the pairs of corpora we built: only one out of two interlinked pages can be kept in one of our corpora. This reduces the comparability of our datasets. However, the two sides of each dataset still share several dimensions along which they are comparable:

- They belong to the same genre distribution, mainly 'encyclopedic article' (see Table 1).
- They were written in the same time period: contemporary prose.
- Because Wikipedia has a dense coverage of many topics, removing one page does not suppress a topic entirely.

As we discuss later, some of the participant systems did detect original sentence pairs that were translations of each other, i.e., that we had not artificially inserted into the monolingual corpora. This is another clue of the comparability of these corpora.

2.2.5. Preparing Training and Test Splits

A shared task dataset needs to have separate training and test splits. Because of the way we selected insertion points for parallel sentences in our initial monolingual corpora, the two sentences of a parallel pair may occur in quite different

[3] We used the following Solr parameters: efType=edismax, qs=5, ps=5, ps2=5, mm=70%, stopwords=true.

[4] https://github.com/attardi/wikiextractor

Pair	Sample (2%)			Training (49%)			Test (49%)		
	fr	en	gold	*fr*	en	gold	*fr*	en	gold
de-en	32593	40354	1038	413869	399337	9580	413884	396534	9550
fr-en	21497	38069	929	271874	369810	9086	276833	373459	9043
ru-en	45459	72766	2374	460853	558401	14435	457327	566356	14330
zh-en	8624	13589	257	94637	88860	1899	91824	90037	1896

Table 2: Corpus statistics: number of monolingual sentences (*fr*, en) and of parallel pairs (gold) for each split and each language pair. The *fr* column stands for the non-English language in each pair. Reprinted from (Zweigenbaum et al., 2017).

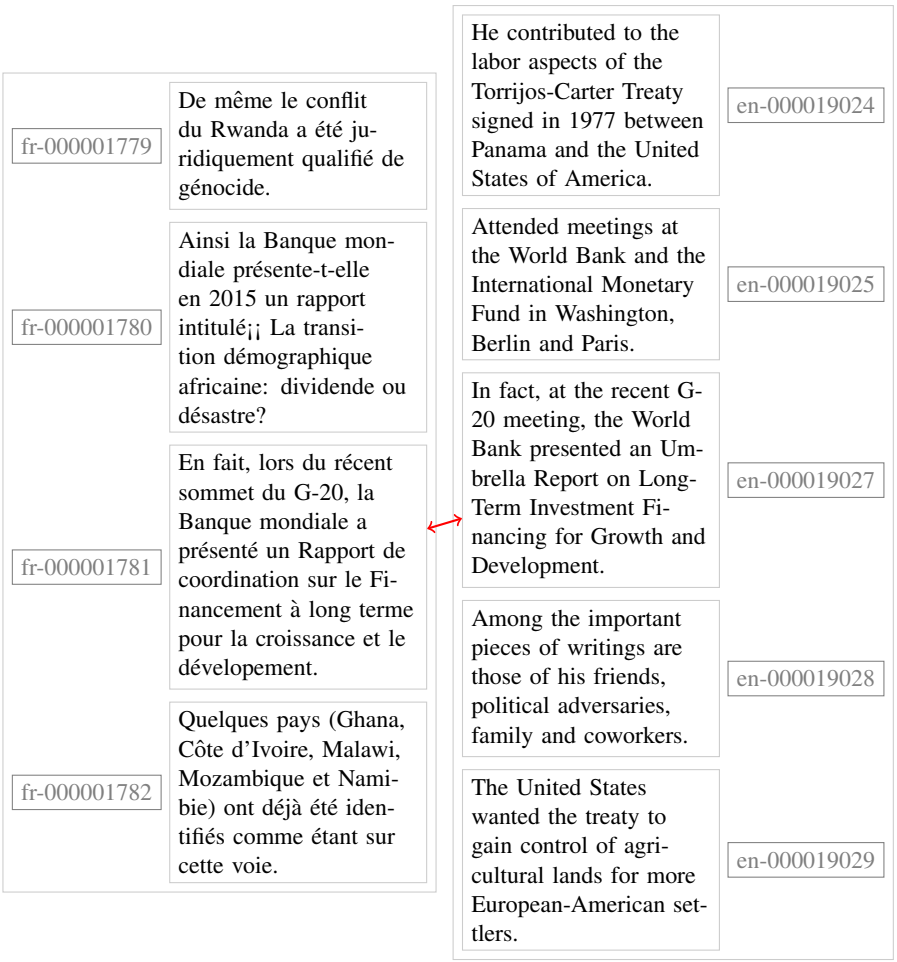

Figure 1: Excerpt from the English-French corpus: fr-000001781 and en-000019027 are inserted parallel sentences.

parts of these monolingual corpora. Splitting the resulting corpora after insertion was therefore liable to separate a large proportion of sentence pairs. Therefore we randomly split the documents of the corpora before parallel sentence pair insertion. An additional constraint was the need to separate interlinked articles. This constraint was taken into account at splitting time, correcting the random assignment of a document to a given split (and redrawing another assignment) if an interlinked document was already present in this split. We actually split the corpora into three parts: a small sample split shown on the shared task Web site, and training and test splits that required registration prior to download. We arbitrary chose 2% of the total for the size of the sample split, and gave training and test half of the remaining data (i.e., 49% each). The corpus preparation process was then performed on each split.[5]

We applied this process to five languages: Chinese (zh), English (en), French (fr), German (de), Russian (ru); this produced four bilingual datasets (see Table 2). Figure 1 shows an example drawn from the English-French dataset.

2.3. The Case of Chinese

Adding a language to our corpora, i.e., providing an additional language paired with English, requires the following data and components:

- A monolingual corpus: Wikipedia is a possible candidate for a large number of languages.
- A parallel corpus with English on one side: in the

[5] An anonymous reviewer rightly pointed out that a development split could have been provided too.

present work we used News Commentary 2016, which pairs English with eleven languages: Arabic, Dutch, Chinese, Czech, German, Spanish, French, Italian, Japanese, Portuguese, and Russian.

- A configuration for indexing and search in the Solr search engine, typically based on a tokenizer, stop words, and possibly more language components.
- Constraints on the range of sentence lengths.

We report here how we included Chinese data in the present corpus.

The Chinese writing system does not separate words with spaces[6]. This raises issues for tokenization that have consequences on our dataset construction pipeline. Various methods have been proposed to tokenize Chinese, including Conditional Random Fields classifiers in the Stanford Chinese Word Segmenter (Tseng et al., 2005) and in the Chinese Mecab[7]. Independently of these methods, several guidelines have been proposed for human annotation of Chinese tokens, including the Chinese Penn Treebank and the Peking University standard (Duan et al., 2003). This results in tokens with shorter or larger spans depending on the guideline, for instance 有线 (*cable*) 电视 (*television*) according to Peking University vs. 有线电视 (*cable television*) according to Chinese Penn Treebank. Chinese tokenizers display the same variety in their choices of token span length; some, such as Stanford or jieba,[8] leave it to the user to choose which strategy to apply (full=short, default=large, search=multiple solutions). We attempted to avoid these considerations by working directly with characters. This was initially motivated by the technical choice of Solr (v6.4.0), whose only option for a Chinese tokenizer was bigrams. We kept sentences between 15 and 40 characters, which we estimated to yield sizes comparable to the English sentences (between 10 and 20 words). Query sentences were truncated to 15 characters (instead of 5 words). However, working with characters raised the following issues. First, character-based Solr search resulted in a lower sentence similarity than in other languages. It often occurred that given a (parallel) Chinese sentence as a query, the matching unigrams or bigrams of characters would be stop words or other common (bigrams of) characters. We compiled a set of 533 stop expressions including punctuation, short common words (一 one, 一切 every), and locutions such as 不仅 (not only), 一方面 (on the one hand), 另一方面 (on the other hand), 反过来说 (on the other hand), etc. Stop expressions were removed from parallel sentences before using them as queries.

Besides, sentences would sometimes start with a common locution followed by a comma: 从历史上看， (from a historical point of view), 由此可见， (from this, it can be seen that), etc. Seeing the large variety of such locutions, we decided to remove any leading sequence of up to six characters followed by a comma from the start of Chinese parallel sentences before using them as queries.

These heuristics were designed and tuned by human review of samples of resulting sequences of two sentences.

3. Use in Two Shared Tasks

These datasets were used in the BUCC 2017 and 2018 Shared Tasks (Zweigenbaum et al., 2017; Zweigenbaum et al., 2018). Participants were taskeed with detecting in a bilingual pair of corpora the inserted parallel sentences. Three of the four language pairs were addressed by the participants in 2017: French, German, and Chinese, with a maximum F-score of 0.84 on German-English (Azpeitia et al., 2017) (see Table 3). All four language pairs were addressed in 2018, with improved F-scores topping at 0.86 for German-English again.

Year	de-en	fr-en	ru-en	zh-en
2017	84	79	–	43
2018	86	81	81	75

Table 3: Best F-scores (%) at the BUCC Shared Tasks in 2017 and 2018

4. Discussion and Perspectives

The resulting corpora can be obtained from the BUCC Web site.[9] They total about 3.5 million sentences in five languages. Participants found methods to cope with this large number of sentences without metadata. To our knowledge, no participant tried to take advantage of a possible lack of cohesion or other features of the inserted sentences that would come from their artificial insertion into the monolingual corpora. The remaining question is that of the accuracy of the provided gold standard, which only accounts for artificially inserted parallel sentences: this accuracy may be reduced by the possible existence of naturally occurring parallel sentence pairs. We estimated their rate of occurrence to be at most 5%, based on a human assessment of the proportion of sentence pairs among the false positives of the most precise systems that are actually true parallel sentence pairs (Zweigenbaum et al., 2017).

5. Acknowledgements

PZ acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207. We thank Zheng Zhang for his help in evaluating the cohesion of Chinese sentence pairs in earlier experiments. Issues in the cohesion of the final corpora are our own responsibility.

6. Bibliographical References

Abdul-Rauf, S. and Schwenk, H. (2009). Exploiting comparable corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora*, BUCC '09, pages 46–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

[6] Throughout this paper we use the term *Chinese* to refer to Modern Standard Chinese, often called Mandarin Chinese.

[7] https://github.com/panyang/MeCab-Chinese

[8] https://github.com/fxsjy/jieba

[9] https://comparable.limsi.fr/bucc2017/bucc2017-task.html

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June. Association for Computational Linguistics.

Azpeitia, A., Etchegoyhen, T., and Martínez Garcia, E. (2017). Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45, Vancouver, Canada, August. Association for Computational Linguistics.

Buck, C. and Koehn, P. (2016). Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany, August. Association for Computational Linguistics.

Duan, H., Bai, X., Chang, B., and Yu, S. (2003). Chinese word segmentation at Peking University. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 152–155, Sapporo, Japan, July. Association for Computational Linguistics.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In Daniel Marcu Susan Dumais et al., editors, *HLT-NAACL 2004: Main Proceedings*, pages 265–272, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In Pamela Forner, et al., editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sharoff, S., Zweigenbaum, P., and Rapp, R. (2015). BUCC shared task: Cross-language document similarity. In *Proc Eighth Workshop on Building and Using Comparable Corpora*, pages 74–78, Beijing, China, July. Association for Computational Linguistics.

Sharoff, S. (2013). Measuring the distance between comparable corpora between languages. In Serge Sharoff, et al., editors, *BUCC: Building and Using Comparable Corpora*, pages 113–129. Springer.

Sharoff, S. (2018). Functional text dimensions for annotation of web corpora. *Corpora*, 13(1).

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411. Association for Computational Linguistics.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A Conditional Random Field word segmenter for SIGHAN Bakeoff 2005. In *SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics, January.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada, August. Association for Computational Linguistics.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2018). Overview of the 2018 BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May. ELRA.