A Parallel Corpus of Arabic–Japanese News Articles

Go Inoue,[1,2] **Nizar Habash,**[2] **Yuji Matsumoto,**[1] **Hiroyuki Aoyama**[3]
[1]Computational Linguistics Laboratory, Nara Institute of Science and Technology, Japan
[2]Computational Approaches to Modeling Language Lab, New York University Abu Dhabi, UAE
[3]Graduate School of Global Studies, Tokyo University of Foreign Studies, Japan
{inoue.go.ib4, matsu}@is.naist.jp, nizar.habash@nyu.edu, aljabal@tufs.ac.jp

Abstract

Much work has been done on machine translation between major language pairs including Arabic–English and English–Japanese thanks to the availability of large-scale parallel corpora with manually verified subsets of parallel sentences. However, there has been little research conducted on the Arabic–Japanese language pair due to its parallel-data scarcity, despite being a good example of interestingly contrasting differences in typology. In this paper, we describe the creation process and statistics of the Arabic–Japanese portion of the TUFS Media Corpus, a parallel corpus of translated news articles collected at Tokyo University of Foreign Studies (TUFS). Part of the corpus is manually aligned at the sentence level for development and testing. The corpus is provided in two formats: A document-level parallel corpus in XML format, and a sentence-level parallel corpus in plain text format. We also report the first results of Arabic–Japanese phrase-based machine translation trained on our corpus.

Keywords: Arabic, Japanese, Parallel Corpus, Sentence Alignment, Machine Translation

1. Introduction

Machine translation (MT) has been a very active research area in natural language processing. Whether its paradigm is statistical or neural, the availability of parallel data is essential for building high-quality systems. In particular, manually verified data sets for development and testing are of great importance for improving and evaluating MT systems. Much work has been done on MT between major language pairs including Arabic–English and Japanese–English, thanks to the availability of large-scale parallel corpora across various domains with manually aligned subsets. However, there has been little research conducted on the Arabic–Japanese language pair due to its parallel-data scarcity, despite being a good example of interestingly contrasting differences in typology. For instance, Arabic is a verb-initial language, while Japanese is a verb-final language, where the position of verb is completely opposite as shown in Figure 1. An Arabic token can be highly ambiguous in morphological, syntactical, and lexical levels due the absence of optional diacritics for short vowels and consonant doubling. In addition, Arabic has a complex system of derivation, inflection, and cliticization. In contrast, a Japanese token can be highly ambiguous due to the absence of spaces between tokens. For more details in linguistic issues, see Habash (2010) for Arabic and Bond and Baldwin (2016) for Japanese.

In this paper, we present a parallel corpus of Arabic–Japanese news articles, part of which is manually aligned at the sentence level for tuning and evaluation. We also provide the first results of Arabic–Japanese phrase-based MT trained on our corpus.

The corpus represents an ongoing project carried out at Tokyo University of Foreign Studies (TUFS) entitled TUFS Media Project,[1] which produces translated news articles in eight languages (Arabic, Bengali, Burmese, Indonesian, Persian, Turkish, Urdu, and Vietnamese). Our corpus serves as a pilot corpus for building parallel corpora of under-resourced language pairs under this project, as well as a basis for investigating various MT techniques for under-resourced language pairs such as pivoting and domain adaptation in future work.

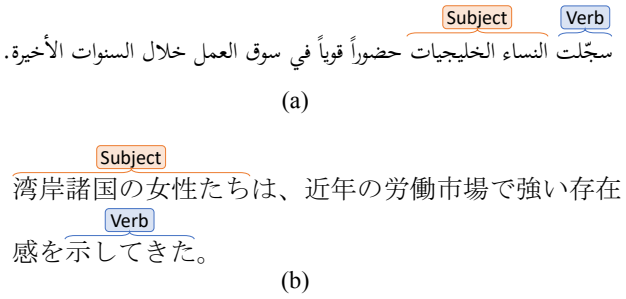

Figure 1: Example of Arabic (a) and Japanese (b) text in parallel. Note that Arabic is written from right to left, and Japanese is written from left to right. *"Women in Gulf countries have shown a strong presence in the recent labour market."*

The corpus is provided in two formats: (a) A document-level parallel corpus of 8,652 document pairs with genre annotation in XML, and (b) a sentence-level parallel corpus in plain text format. The sentence-level parallel corpus consists of 64,488 sentence pairs, with approximately 2.4 million Arabic tokens[2] and 3.7 million Japanese tokens[3] in total. Our corpus is publicly available for research purposes.[4]

[1]http://www.el.tufs.ac.jp/tufsmedia/

[2]Throughout this paper, an Arabic token is defined as a simple tokenization unit (D0) (Habash, 2010) as shown in Table 1.

[3]A Japanese token is defined as a unit used in IPAdic (2.7.0.) (Asahara and Matsumoto, 2003).

[4]http://el.tufs.ac.jp/tufsmedia-corpus/

Tokenization	Operation	Example
raw	no tokenization	*wsyktbhA llTAlb.*
D0	split punctuations and numbers	*wsyktbhA llTAlb* ␣.
D1	split CONJ	*w+* ␣ *syktbhA llTAlb* ␣.
D2	split CONJ and PART	*w+* ␣ *s+* ␣ *yktbhA l+* ␣ *AlTAlb* ␣.
ATB	split all clitics except the definite article	*w+* ␣ *s+* ␣ *yktb* ␣ *+hA l+* ␣ *AlTAlb* ␣.
D3	split all clitics	*w+* ␣ *s+* ␣ *yktb* ␣ *+hA l+* ␣ *Al+* ␣ *TAlb* ␣.
D3*	remove the definite article from D3	*w+* ␣ *s+* ␣ *yktb* ␣ *+hA l+* ␣ *TAlb* ␣.

Table 1: Examples of the various tokenization schemes for the raw input وسيكتبها للطالب. *wsyktbhA llTAlb.* 'And he will write it for the student.' Arabic characters are transliterated in the Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007). The symbol "␣" denotes a space added after tokenization. CONJ and PART refer to conjunctions and particles, respectively.

2. TUFS Media Corpus

In this section, we describe the source of our corpus, details of the corpus construction process, and statistics of our corpus.

2.1. Source of the Corpus

TUFS Media Project is an ongoing project carried out at Tokyo University of Foreign Studies to offer translated news from various countries and regions around the world in order to familiarize the Japanese society with the current world events. The initial version of the project was launched in 2005, offering translated articles from three languages: Arabic, Turkish, and Persian. Currently, the project provides translated articles into Japanese from eight languages, Arabic, Bengali, Burmese, Indonesian, Persian, Turkish, Urdu, and Vietnamese.

Translation of an article is done in two steps, initial translation and proofreading. In the initial translation step, a translator, typically an undergraduate student who majors in Arabic, choses an appropriate article from one of the news agencies[5] in accordance with the person's interest following the translation guideline. The guideline describes rules regarding the choice of an article to be translated, formatting, and transcription. The translator then translates title, dateline, and paragraphs in the article. The paragraphs can be omitted as long as the text to be translated contains over 200 words in Arabic. In that case, the translator inserts a phrase that denotes omission in the translated article. The translator also assigns a concise title and classifies the article into 12 categories based on the content. In the proofreading step, a proofreader, who is an expert in Arabic or a graduate student with experience studying in the region, proofreads the translated article and publishes it on the project website.

2.2. Corpus Construction

We describe next the process of corpus construction, from data collection to sentence alignment.

2.2.1. Crawling Documents from Project Website

We crawled the project website[6] which provides a searchable interface for translated articles, specifying the six news agencies and the issue date that ranges from 2005 to 2016. The crawling yielded 9,915 translated articles in HTML format. For Arabic, we collected original articles by crawling the provided links to the original urls and the archived versions in MHTML format. MHTML files were converted to HTML files in order to simplify the succeeding scraping process. The crawling of original articles yielded 9,056 documents[7] in total.

2.2.2. Scraping Crawled Documents

For Japanese, we extracted translated text, category, issue date, and links to the original articles from the documents using HTML tags as clues. The Japanese data are more structured than the Arabic data thanks to the unified HTML architecture and the translation guidelines, however, there are some cases where we could not find corresponding translations for the original title and/or dateline.

The Arabic data are more difficult to process due to their format variations across six different agencies with periodically different templates within agencies. In some cases, we could not extract main texts from the documents due to their structural issues in HTML. In such cases, we simply discarded these documents from our corpus. Paragraph boundaries are kept in both languages.

2.2.3. Text Cleaning and Formatting

We identified and removed any notes translators may have made, in order to keep the parallel texts as comparable as possible. We also deleted documents that are not detected as Arabic contents by a python library `langdetect` (1.0.7).[8] Finally, we took the intersection of the parallel documents in both languages, yielding 8,652 document pairs. All documents are segmented into sentences by `Pragmatic Segmenter` (0.3.16),[9] a rule-based sentence splitter. For Arabic, we used full stop, exclamation mark, and question mark for the set of delimiters. For Japanese, we used the default set of delimiters defined in the segmenter. The document-level aligned corpus is available in XML with UTF-8 encodings as shown in Figure 2.

[5]The agencies are: Al-Ahram, Al-Hayat, Al-Nahar, Al-Quds Al-Arabi, Al-Sabah Al-Jadid, and Al-Watan.

[6]`http://www.el.tufs.ac.jp/prmeis/`

[7]The decrease in the number of Arabic documents is due to the absence of valid links to the original ones or conversion error from MHTML to HTML.

[8]`https://pypi.python.org/pypi/langdetect/`

[9]`https://github.com/diasks2/pragmatic_segmenter/`

<?xml version="1.0" encoding="UTF-8" ?>
<body>
<meta>
<article_id>News20120225_100246</article_id>
<agency>Al-Hayat</agency>
<lang>Arabic</lang>
<category>Economy</category>
</meta>
<content>
<title>
<t id="1">500 بليون دولار ثروات النساء العربيات</t>
</title>
<dateline>دبي - دلال أبو غزالة</dateline>
<text>

<s id="1:1">سجّلت النساء الخليجيات حضوراً قوياً في سوق العمل خلال السنوات الأخيرة، وزاد عددهن إلى 3.3 مليون، مقارنة بـ1.5 مليون في العقد الأول من القرن الـ21، بزيادة 83 في المئة.</s>

<s id="2:1">ووفقاً لـ «مجموعة بوسطن للاستشارات» فإن حجم الثروة المتركزة لدى النساء في المنطقة العربية تصل إلى 500 بليون دولار، بينما تقدر مجلة «ميد» حجم الثروة لدى النساء الخليجيات بـ385 بليوناً.</s>

<s id="3:1">وأكدت مؤسسة «الماسة كابيتال» في تقرير صدر أمس، أن ما يزيد على 26 في المئة من النساء في المنطقة العربية دخلن إلى سوق العمل، علماً أن النساء يشكلن 41 في المئة من تعداد السكان.</s>

<s id="4:1">وتستثمر هذه الثروات عادة في الأصول الآمنة كالسندات والودائع النقدية والمصرفية، ما دفع بعض المؤسسات المالية إلى اتخاذ خطوات جدية للاستفادة من هذا النهج.</s>
<s id="4:2">وكرست المؤسسات المالية هذه الأموال والاستفادة منها من خلال تأسيس فروع مصرفية خاصة بالسيدات، وتأسيس صناديق موجّهة إليهن.</s>

</text>
</content>
</body>

<?xml version="1.0" encoding="UTF-8" ?>
<body>
<meta>
<article_id>News20120225_100246</article_id>
<agency>Al-Hayat</agency>
<lang>Japanese</lang>
<category>Economy</category>
</meta>
<content>
<title>
<t id="1">■アラブ人女性の総資産、5000億ドルへ</t>
</title>
<dateline>ドバイ：ダラール・アブー・ガザーラ
</dateline>
<text>

<s id="1:1">湾岸諸国の女性たちは、近年の労働市場で強い存在感を示してきた。</s>
<s id="1:2">労働者数は、21世紀初めの10年では150万人だったのに比べ、83%増の330万人に増加した。</s>

<s id="2:1">「ボストンコンサルタントグループ」によると、アラブ地域の女性たちが保有する資産規模は、5,000億ドルに達する。</s>
<s id="2:2">一方『ミード』誌は、湾岸諸国の女性たちが保有する資産規模を3850億ドルと推計する。</s>

<omit>（後略）</omit>
</text>
</content>
</body>

Figure 2: An example of a parallel document pair in XML format.

2.2.4. Manual Sentence Alignment for Evaluation

We manually aligned the latest 900 documents in publication date for evaluation purposes. We divided 900 documents into three divisions for blind-test, dev-test, and dev-tune sets. The divisions are as follows: The latest 400 documents for the blind-test set, the second latest 400 documents for the dev-test set, and the third latest 100 documents for the dev-tune set. We used the InterText tool (Vondřička, 2014) to create alignment files.

Arabic-to-Japanese	Sentence Pairs	Percentage
1-to-0	7,624	58.75
1-to-1	2,758	21.25
1-to-many	2,328	17.94
0-to-1	102	0.79
many-to-1	83	0.64
many-to-many	81	0.62

Table 2: Types of sentence alignment pairs in manually aligned data set of 900 documents.

Table 2 shows the distribution of types of sentence alignment pairs in our manually aligned data set. The possible combinations of sentence alignment pairs are as follows: One sentence in one language corresponds to one sentence in another (1-to-1), one sentence does not have corresponding sentence (1-to-0, 0-to-1), one sentence corresponds to multiple sentences (1-to-many, many-to-1), and multiple sentences correspond to multiple sentences (many-to-many). Three documents were not aligned in the document level due to the modification in the original article after translation.

The large number of 1-to-0 alignments is due to the extra paragraphs in the Arabic side that are not translated into Japanese. Apart from the null alignments (1-to-0, 0-to-1), 1-to-1 alignments account for 52.53%, whereas 1-to-many ones account for 44.34%. This can be attributed to the difference between Arabic and Japanese in punctuation usage and stylistic preference in the translation process.

2.2.5. Automatic Sentence Alignment

We compare three different alignment tools, a python implementation (Tan and Bond, 2014) of the algorithm of Gale and Church (1993), HunAlign (Varga et al., 2005), and Gargantua (Braune and Fraser, 2010).

Preprocessing We lemmatized both Arabic and Japanese texts before running a sentence aligner. We used the MADAMIRA toolkit (Pasha et al., 2014) for Arabic and MeCab (0.996) (Kudo, 2005) with IPAdic for Japanese. We performed NFKC normalization before lemmatizing Japanese tokens.

We deleted untranslated paragraphs in the Arabic side so that the number of paragraphs should be the same in both documents. This process is done only for the documents with an explicit markup that denotes omission in the latter part of a Japanese document. We used paragraph boundaries as a hard delimiter.

HunAlign To employ HunAlign, we use an approach similar to the three-step workflow used in the JRC-Arcquis corpus (Steinberger et al., 2016) and DCEP corpus (Hajlaoui et al., 2014), which consists of an initial alignment using length similarity, automatic dictionary construction from the initial alignments, and a second alignment using lexical similarity calculated with the constructed dictionary in the second step. Specifically, we first run HunAlign to obtain the initial alignment without dictionary, randomly sample 10,000 sentence pairs from the 1-to-1 segments in the initial alignment, build a dictionary with minimum occurrence score of 2 and minimum association score of 0.2, and finally, re-align all sentences with the constructed dictionary.

Evalutation We evaluate the quality of sentence alignment using the dev-tune set. We measure precision, recall, and F_1 scores in the sentence level. Precision is defined as the ratio of the number of correctly aligned pairs divided by the number of predicted pairs. Recall is defined as the ratio of the number of correctly aligned pairs divided by the number of reference pairs. F_1 is defined as the harmonic mean of precision and recall.

Results Table 3 shows the performance of the three alignment algorithms. The low F_1 score of the Gale and Church (1993) algorithm can be attributed to the distribution of alignment types and the imperfect alignment in the paragraph level. HunAlign and Gargantua outperformed the Gale and Church (1993) algorithm by large and yielded comparable results.

Alignment Algorithm	P	R	F_1
Gale and Church (1993)	0.49	0.45	0.47
HunAlign	0.74	0.77	0.76
Gargantua	**0.76**	**0.80**	**0.78**

Table 3: Sentence alignment precision (P), recall (R), and F_1 scores on dev-tune set.

2.3. Corpus Statistics

In Table 4, we provide the distribution of the categories in our corpus as determined by translators. Table 5 shows the basic statistics of sentence-level parallel corpus, including manually aligned sentences and automatically aligned sentences using Gargantua. A large difference in the number of tokens can be attributed to the difference in their tokenization schemes. We segment Japanese tokens in a more fine-grained manner than we segment Arabic tokens. In Modern Standard Arabic, an orthographically single token can have up to four syntactically independent clitics around the stem. If we were to impose D3 tokenization on the Arabic, a scheme which separates all clitics, then we would have a unit much more comparable to Japanese tokens. We ran MADAMIRA on our corpus to obtain the number of D3-tokenized tokens, which was approximately 3.4 million. This is much closer to the number of Japanese tokens, 3.7 million.

Category	Documents	Percentage
Politics	4,253	49.16
International	1,854	21.43
Society	811	9.37
Economy	608	7.03
Column	330	3.81
Lebanon Issue	280	3.24
Culture	244	2.82
Accident	194	2.24
Sports	48	0.55
Others	15	0.17
Nuclear Issue	10	0.12
Book Introduction	5	0.06
Total	8,652	100.00

Table 4: Category distribution of our entire corpus.

3. Machine Translation Baselines

In this section, we present the baseline results of phrase-based MT from Arabic to Japanese.

3.1. Experimental Settings

Phrase-based MT Settings We use the Moses toolkit (Koehn et al., 2007) to build a standard phrase-based MT system. Word alignment was extracted by MGIZA++ (Gao and Vogel, 2008) with a maximum phrase size of 8. We use the *grow-diag-final-and* and *msd-bidirectional-fe* options for symmetrization and reordering. We train a 5-gram language model on the target side of the training set using KenLM (Heafield, 2011). We use MERT (Och, 2003) for decoding weight optimization.

Data and Preprocessing We use the manually aligned data described in Section 2.2.4. for tuning and testing, and the automatically aligned data using Gargantua described in Section 2.2.5. for training.

We tokenize Arabic data using the MADAMIRA toolkit (Pasha et al., 2014) with six tokenization schemes (D0, D1, D2, D3, D3*, and ATB) following Zalmout and Habash (2017). Examples of the six tokenization schemes are shown in Table 1.

We normalize Japanese texts using the NFKC normalization and tokenize them using the MeCab morphological analyzer (0.996) (Kudo, 2005) with IPAdic.

We eliminate long sentences with more than 100 words using the script `clean-corpus-n.perl` before training translation models. Table 6 shows statistics of training data after cleaning.

Evaluation Before evaluating, we de-tokenize the predicted output by deleting spaces between Japanese characters, and then re-tokenize them using MeCab with IPAdic. We calculate automatic evaluation scores for two metrics: BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010). We use the `multi-bleu.perl` script in the Moses toolkit to compute BLEU scores. We calculate RIBES scores using the `RIBES.py` (1.03.1.).[10]

[10] `http://www.kecl.ntt.co.jp/icl/lirg/ribes/`

	Documents	Sentences	Tokens (ar)	Tokens (ja)
dev-tune	100	621	23,312	36,595
dev-test	400	2,393	92,760	147,536
blind-test	400	2,236	85,940	144,358
train	7,752	59,238	2,175,438	3,403,244
Total	8,652	64,488	2,377,460	3,731,733

Table 5: The basic statistics of our parallel corpus. Sentences in the training set are aligned using Gargantua.

Tokenization	Sentences	Tokens (ar)	Tokens (ja)
D0	54,223	1,811,540	2,792,333
D1	54,123	1,939,953	2,785,051
D2	54,029	2,027,916	2,778,315
ATB	53,933	2,107,024	2,771,255
D3	53,164	2,467,747	2,715,485
D3*	53,933	2,107,024	2,771,255

Table 6: The statistics of cleaned corpus for training translation models.

Results Table 7 summarizes the baseline results of phrase-based MT systems across six different tokenization schemes. The D3* scheme performs the best in both BLEU and RIBES scores, followed by the ATB scheme. The result is consistent with Zalmout and Habash (2017), where they show that removing the definite article (ال *Al* 'the') in the Arabic side enhances the performance when translating into the languages without its clear equivalent, Russian and Chinese in their case. This result is understandable since Japanese also lacks the definite article.

	dev-test		blind-test	
Tokenization	BLEU	RIBES	BLEU	RIBES
D0	10.78	56.61	8.76	55.71
D1	10.70	56.90	8.83	55.63
D2	11.13	56.94	9.34	56.08
ATB	11.29	57.54	9.24	56.41
D3	10.53	56.80	8.56	55.77
D3*	**11.48**	**57.86**	**9.38**	**56.63**

Table 7: BLEU and RIBES scores of Arabic–Japanese PBMT systems with different tokenization schemes in the source side.

4. Related Work

Much work has been done on building multilingual parallel corpora which include the language pair of Arabic and Japanese. Table 8 summarizes the statistics of publicly available parallel corpora of this language pair. Lison and Tiedemann (2016) presents the largest corpus, in which they collected movie and TV subtitles from OpenSubtitles.[11] Cettolo and Girardi (2012) constructed a parallel corpus that consists of transcribed and translated TED talks. Abdelali et al. (2014) developed the AMARA corpus that includes subtitles of educational video lectures on Massive Online Open Courses (MOOCs). Christodouloupoulos and Steedman (2015) presents a collection of Bible translations across 100 languages. Tiedemann (2012) provides a collection of Quran translations (Tanzil), localization files of technical manuals (GNOME, Ubuntu, and KDE4), as well as the collections of translations in the news domain (Global Voices, Tatoeba, News-Commentary 11). Prokopidis et al. (2016) constructed parallel corpora from Global Voices similar to Tiedemann (2012).

Compared to the domains such as subtitles, religious texts, and technical manuals, the amount of data in the news domain is very limited. Our corpus aims to supplement the lack of parallel data in this domain by constructing a parallel corpus with over 64,000 sentences (2.4 million Arabic tokens and 3.7 million Japanese tokens), including manually aligned sentence pairs for development and evaluation.

5. Conclusion and Future Work

We presented a parallel corpus of Arabic–Japanese news articles comprising 8,652 document pairs. Part of the corpus is manually aligned at the sentence level for development and testing. The corpus is provided in two formats: (a) A document-level parallel corpus with genre annotation in XML, and (b) a sentence-level parallel corpus in plain text format. The sentence-level parallel corpus comprises 64,488 sentence pairs with approximately 2.4 million Arabic tokens and 3.7 million Japanese tokens. We also reported the first results of Arabic–Japanese phrase-based MT trained on our corpus.

As future work, we will explore sentence alignment methods to improve the quality of our corpus. We also plan to explore MT techniques for under-resourced language pairs such as pivoting, and domain adaptation from better resourced domains.

6. Acknowledgements

This work has been partially funded by the Tobitate! (Leap for Tomorrow) Young Ambassador Program. Part of the work was done during the first author's visit to New York University Abu Dhabi. The creation of translated articles has been carried out at Tokyo University of Foreign Studies. We are grateful to the contributors of the TUFS Media Project for providing translated articles. We thank the anonymous reviewers, Nasser Zalmout, Alexander Erdmann, Salam Khalifa, and Ossama Obeid for their helpful comments.

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

	Sentences	Tokens (ar)	Tokens (ja)	Domain
OpenSubtitiles2018 (Lison and Tiedemann, 2016)	1,834,940	11,615,534	13,319,298	Subtitle
TED (Cettolo and Girardi, 2012)	205,734	1,426,132	1,857,188	Subtitle
AMARA (Abdelali et al., 2014)	46,457	334,890	486,229	Subtitle
Bible (Christodouloupoulos and Steedman, 2015)	31,067	473,002	1,107,641	Religious
Tanzil (Tiedemann, 2012)	12,471	526,469	526,913	Religious
KDE4 (Tiedemann, 2012)	100,967	552,178	931,438	Technical Manual
Ubuntu (Tiedemann, 2012)	740	4,152	6,272	Technical Manual
GNOME (Tiedemann, 2012)	450	1,247	1,381	Technical Manual
Global Voices (Tiedemann, 2012)	4,929	85,961	121,234	News
Global Voices (Prokopidis et al., 2016)	7,211	127,737	200,215	News
Tatoeba (Tiedemann, 2012)	1,134	6,039	10,947	News
News-Commentary11 (Tiedemann, 2012)	569	39,937	52,085	News
Our Corpus	64,488	2,377,460	3,731,733	News

Table 8: Statistics of publicly available parallel corpora of Arabic and Japanese.

7. Bibliographical References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Asahara, M. and Matsumoto, Y., (2003). *ipadic version 2.7.0 User's Manual (in Japanese)*. Nara Institute of Science and Technology.

Bond, F. and Baldwin, T. (2016). Introduction to Japanese Computational Linguistics. In Kentaro Inui Shun Ishizaki Hiroshi Nakagawa Francis Bond, Timothy Baldwin et al., editors, *Readings in Japanese Natural Language Processing*, CSLI Studies in Computational Linguistics, pages 1–28. CSLI Publications.

Braune, F. and Fraser, A. (2010). Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 81–89. Association for Computational Linguistics.

Cettolo, M. and Girardi, C. (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395, Jun.

Gale, W. A. and Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational linguistics*, 19(1):75–102.

Gao, Q. and Vogel, S. (2008). Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*, volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). DCEP – Digital Corpus of the European Parliament. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA, October. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Kudo, T. (2005). MeCab: Yet Another Part-of-Speech and Morphological Analyzer. *http://mecab. sourceforge. net/*.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, volume 14, pages 1094–1101, Reykjavik, Iceland.

Prokopidis, P., Papavassiliou, V., and Piperidis, S. (2016). Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2016). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, may.

Tan, L. and Bond, F. (2014). NTU-MC Toolkit: Annotating a Linguistically Diverse Corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, volume 2012, pages 2214–2218, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2005). Parallel Corpora for Medium Density Languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Vondřička, P. (2014). Aligning Parallel Texts with InterText. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Zalmout, N. and Habash, N. (2017). Optimizing Tokenization Choice for Machine Translation across Multiple Target Languages. *The Prague Bulletin of Mathematical Linguistics*, 108(1):257–269.