A Parallel Corpus for Evaluating Machine Translation between Arabic and European Languages

Nizar Habash, Nasser Zalmout, Dima Taji, Hieu Hoang and Maverick Alzate
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
{nizar.habash,nasser.zalmout,dima.taji,hh65,ma2835}@nyu.edu

Abstract

We present Arab-Acquis, a large publicly available dataset for evaluating machine translation between 22 European languages and Arabic. Arab-Acquis consists of over 12,000 sentences from the JRC-Acquis (Acquis Communautaire) corpus translated twice by professional translators, once from English and once from French, and totaling over 600,000 words. The corpus follows previous data splits in the literature for tuning, development, and testing. We describe the corpus and how it was created. We also present the first benchmarking results on translating to and from Arabic for 22 European languages.

1 Introduction

Statistical Machine Translation (SMT, henceforth MT) is a highly data driven field that relies on parallel language datasets for training, tuning and evaluation. Prime examples of such modern-day digital *Rosetta Stones* include the United Nations corpus (six languages) and the European Parliamentary Proceedings corpus (20+ languages).[1] MT systems use these resources for model development and for evaluation. Large training data is often not available and researchers rely on other methods, such as pivoting to build MT systems. And while this addresses the question of training, there is still a need to tune and evaluate. In the case of Arabic, most of MT research and MT evaluation resources are focused on translation from Arabic into English, with few additional resources pairing Arabic with a half dozen languages. This paper showcases the effort to create a dataset, which we dub **Arab-Acquis**, to support the development and evaluation of machine translation systems from Arabic to the languages of the European Union and vice versa. Our approach is simply to exploit the existence of the **JRC-Acquis** corpus (Steinberger et al., 2006; Koehn et al., 2009), which has 22 languages in parallel, and translate a portion of it to Standard Arabic. We include two translations in Arabic for each sentence in the set to support robust multi-reference evaluation metrics. This provides us with the largest (and first of its kind) set of multilingual translation for Standard Arabic to date. It allows us to evaluate the quality of translating into Arabic from a set of 22 languages, most of which have no large high quality datasets paired with Arabic.

2 Related Work

In the context of MT research in general, multilingual resources (or parallel corpora) are central. Some of these resources exist naturally such as the United Nations corpus (Arabic, Chinese, English, French, Russian and Spanish) (Rafalovitch et al., 2009), the Canadian Hansards (French and English) (Simard et al., 1993), the European Parliament proceedings, EUROPARL, (21 languages in its latest release) (Koehn, 2005), and the **JRC-Acquis** (22 languages) (Steinberger et al., 2006; Koehn et al., 2009). Translations may also be commissioned to support MT research, as in the creation of an Arabic dialect to English translation corpus using crowdsourcing (Zbib et al., 2012). Such resources are necessary for the development of MT systems, and for the evaluation of MT systems in general. While training MT systems typically requires large collections in the order of millions of words, the automatic evaluation of MT requires less data; but evaluation data is expected to

[1] The European Parliament has 24 official languages (European Parliament, 2016); however the corpus we used only contained 22, missing only Irish and Croatian (Steinberger et al., 2006; Koehn et al., 2009).

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 235–241,
Valencia, Spain, April 3-7, 2017.

have more than one human reference since there are many ways to translate from one language to another (Papineni et al., 2002). The number of language pairs that are fortunate to have large parallel data is limited. Researchers have explored ways to exploit existing resources by pivoting or bridging on a third language (Utiyama and Isahara, 2007; Habash and Hu, 2009; El Kholy et al., 2013). These techniques have shown promise but can obviously only be pursued for languages with parallel evaluation datasets, which are not common. In some cases, researchers translated commonly used test sets to other languages to enrich the parallelism of the data, e.g., (Cettolo et al., 2011), while working on Arabic-Italian MT, translated a NIST MT eval dataset (Arabic to four English references) to French and Italian. For Arabic MT, the past 10 years have witnessed a lot of interest in translating from Arabic to English mostly due to large DARPA programs such as GALE and BOLT (Olive et al., 2011). There have been some limited efforts in comparison on translating into Arabic from English (Hamon and Choukri, 2011; Al-Haj and Lavie, 2012; El Kholy and Habash, 2012), but also between Arabic and other languages (Boudabous et al., 2013; Habash and Hu, 2009; Shilon et al., 2012; Cettolo et al., 2011). The **JRC-Acquis** collection, of which we translate a portion, is publicly available for research purposes and already exists in 22 languages (and others ongoing). As such, the **Arab-Acquis** dataset will open a pathway for researchers to work on MT from a large number of languages into Arabic and vice versa, covering pairs that have not been researched before. The dataset enables us to compare translation quality from different languages into Arabic without data variation. In this paper, we also present some initial benchmarking results using sentence pivoting techniques between all **JRC-Acquis** languages and Arabic.

3 Approach and Development of Arab-Acquis

We discuss next the design choices and the process we followed to create **Arab-Acquis**.

3.1 Desiderata

As part of the process of creating the **Arab-Acquis** translation dataset, we considered the following desiderata:

- The dataset should have a large number of translations to maximize the parallelism.
- The original text should not have any restrictive copyrights.
- It is more desirable to extend datasets and data splits that are already used in the field
- The dataset must be large enough to accommodate decent sized sets for tuning, development, and one or two testing versions.
- Each sentence is translated at least twice, by different translators from different languages.
- It is preferable to use professional translators with quality checks than to use crowdsourcing with lower quality translations.

3.2 Why JRC-Acquis?

Keeping these desiderata in mind, we decided to use the **JRC-Acquis** dataset (Steinberger et al., 2006; Koehn et al., 2009) as the base to select translations from. **JRC-Acquis** is the JRC (Joint Research Centre) Collection of the *Acquis Communautaire*, which is the body of common rights and obligations binding all the Member States together within the European Union (EU). By definition, translations of this document collection are therefore available in all official EU languages (Steinberger et al., 2006). The corpus version we use contains texts in 22 official EU languages (see Table 2). The **JRC-Acquis** corpus text is mostly legal in nature, but since the law and agreements cover most domains of life, the corpus contains vocabulary from a wide range of subjects, e.g., human and veterinary medicine, the environment, agriculture, commerce, transport, energy, and science (Koehn et al., 2009).

The **JRC-Acquis** is also a publicly available dataset that has been heavily used as part of international translation research efforts and shared tasks. It has a lot of momentum that comes from people having worked with. We follow the data split guidelines used by Koehn et al. (2009) and only translate portions that are intended for tuning, development and testing. These portions sum to about 12,000 sentences in total. All mentions of **JRC-Acquis** in the rest of this document will refer to the portion selected for translation into **Arab-Acquis** and not the whole **JRC-Acquis** corpus.

3.3 Translating the JRC-Acquis

For each sentence in **JRC-Acquis**, we created two Arabic references starting with English in one and

French in the other. The choice of these two languages is solely reflective of their prominence in the Arab World. The two languages also have different structures and features that seed differences in wording, which is desirable for such a dataset.

We commissioned three individual companies (from Egypt, Lebanon and Jordan each) to translate the **JRC-Acquis** corpus into Arabic from both English and French. On average, the translation from English cost USD $0.056 per word (for 327,466 words), and the translation from French cost USD $0.073 per word (for 340,739 words). In total the translation cost just over USD $43,200. The files were distributed so that none of the companies would get the same file in both English and French. This allowed for two different translations for each file. The companies took 44 to 90 days to translate the files (65 working days on average).

We instructed the translation companies to maintain the original line formatting. We also stressed that the translation should be in the most natural and fluent Arabic to the translators. We did regular checks on the translations we received from the translation companies, regarding both translation and formatting.

	JRC-Acquis		**Arab-Acquis**	
	English	**French**	**Arabic**$_{En}$	**Arabic**$_{Fr}$
Tune	108,405	112,984	107,271	113,942
Dev	109,611	114,327	114,903	114,795
Test	109,450	113,428	118,491	117,942
Total	327,466	340,739	340,665	346,679

Table 1: **Arab-Acquis** data set sizes, and the sizes of the corresponding sentences (4,108 sentences for *Dev*, 4,107 for rest) in **JRC-Acquis**.

3.4 Arab-Acquis Dataset

In Table 1, we present the final dataset sizes for **Arab-Acquis** and the respective dataset sizes from the **JRC-Acquis** English and French portions used to translate it. In total, we created 687,344 translated words.

4 Translation Analysis

When analyzing the differences in the translations from the English and French sources, we noticed the most variations fall into two categories:

Source Language Bias Since different languages have different styles of writing, these differences are reflected in translations from different language sources (Volansky et al., 2015). An example of such differences includes directive numbers. For example, directives from the *European Economic Community* include the abbreviation *EEC* in English, while in French it becomes *CEE* for *Communauté Économique Européenne*: compare directives *75/34/EEC* (English) and *75/34/CEE* (French).

Valid Alternatives Arabic is a lexically and morphologically rich language; and as such statements can be expressed in different valid styles and sentence structures, and using different alternative wordings that still convey the same meaning. An example of such alternatives is the use of *yly*[2] يلي and *yÂty* يأتي, which are both valid translations for the word '*following*.'

We consider these differences features that make the corpus more suitable to evaluate MT systems by providing more options to express the same concept.

5 Machine Translation Results

In this section we present the first results ever reported on benchmarking MT between Arabic and 22 European languages in both directions using the same datasets and conditions.

5.1 JRC-Acquis MT Systems

We built 21 MT systems for translating from English to X and 21 MT systems for translating from X to English, for X being all of the **JRC-Acquis** languages, other than English. We built these MT systems using the **full JRC-Acquis** corpus following the same data splits for training, tuning, and development used by Koehn et al. (2009), who reported their work on developing 462 machine translation systems based on the 22 languages of the **JRC-Acquis** corpus. Their paper included both direct and pivoting-based systems on multiple languages. We replicated the MT systems in (Koehn and Haddow, 2009), in an effort to pivot from/to Arabic through English. We present the MT results for the European languages with English in Table 2. Our results almost match those at (Koehn et al., 2009). Any minor differences in the scores are mainly attributed to the various upgrades in the toolkits used and tuning variations.

We used the Moses toolkit (Koehn et al., 2007) with default parameters to develop the systems,

[2]Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

along with the extra settings used at the original paper; including limiting the training sentence length to 80 words, and the tuning sentences to 8-60 words long only. We used a 5-gram language model. For systems evaluation, we also use BLEU score (Papineni et al., 2002) through the scripts at Moses. To match the settings used at Koehn's paper, we use the case insensitive evaluation feature of BLEU. We used these settings across all experiments, unless explicitly specified.

5.2 Arabic-English Systems

We used the Arabic-English parallel component of the UN Corpus to train the Ar-En systems. The UN Corpus has a close parliamentary-styled discourse to **JRC-Acquis**'s, which should reduce the divergence with the rest of **JRC-Acquis** MT systems. We used about 9 million lines for the Arabic and English language models (circa 286 million words), 2.4 million parallel lines for training (circa 62 million words) and 2000 lines for tuning. We tokenized the Arabic content using the MADAMIRA toolkit (Pasha et al., 2014) with the Alif/Ya normalized ATB scheme (Habash, 2010), and rule-based detokenization (El Kholy and Habash, 2010) for the resulting translations. The English content was tokenized using the available English tokenizer at Moses. For the translations to Arabic, we used the English and French Arabic translations of the **Arab-Acquis** *Dev* files as two references for BLEU evaluation. For systems translating from Arabic to English, we used only the **Arab-Acquis** Arabic translation from the English sources for our tuning.

We compared the performance on an in-domain data set from the UN Corpus with the performance on the Arabic-English dataset from **Arab-Acquis**. The in-domain results were 43.09 and 39.29 for Ar-En and En-Ar respectively, whereas the out-of-domain scored 28.76 and 27.83. As expected, the performance on in-domain data is much better than on out-of-domain. The out-of-domain results reflect the systems used in the pivoting.

5.3 Pivoting through English

We used the English part of the shared **Arab-Acquis** content for pivoting from Arabic into the remainder of the **JRC-Acquis** languages. This approach can be used to test and validate further pivoting research involving Arabic, with diverse target/input languages. Instead of building MT systems for a given language with Arabic, pivoting can be used as a viable option in many scenarios. We used simple chaining of the source-pivot system and the pivot-target system when translating from/to Arabic and the various **JRC-Acquis** languages, where the pivot language was always English. We leave exploring more sophisticated pivoting techniques (Utiyama and Isahara, 2007; Habash and Hu, 2009; El Kholy et al., 2013) and newer neural machine translation techniques (Johnson et al., 2016) to future work. The results are presented in Table 2.

5.4 Discussion

Table 2 specifies for each language X four BLEU scores for translation from and to English (En→X and X→En), and from and to Arabic via English pivoting (Ar→En→X and X→En→Ar).

Direct English MT Our En→X and X→En results are generally comparable to those reported by Koehn et al. (2009). The highest BLEU score in the En→X direction is for French, and the worst BLEU score is for Hungarian. The highest BLEU score in the X→En direction is for Maltese, and the worst BLEU score is for Hungarian again. This high BLEU score for Maltese is rather surprising, but consistent with (Koehn et al., 2009). Although Maltese is a Semitic language, it has a strong Italian (Romance) component; and English is an official language of the nation of Malta. Also, while Maltese is morphologically rich, its writing system has heavy use of hyphens (e.g., *il-kondizzjonijiet* 'the-conditions') which allows for easy morphological tokenization with simple white space and punctuation tokenization technique used in Moses.

Pivoting through English The BLEU scores for Ar→X and X→Ar via English pivot are to our knowledge the first large scale benchmark of a publicly available data set comparing machine translation from/to Arabic across a large number of languages under identical settings. Not surprisingly, the correlation between the performance on the direct-with-English and pivot-via-English systems is very high: X→En and X→En→Ar correlate at $r = 0.97$, and En→X and Ar→En→X correlate at $r = 0.93$. As such, the highest BLEU score in the Ar→En→X direction is for French again, but the worst BLEU score is for Estonian (a relative of Hungarian from the Finno-Ugric family). The highest BLEU score in the X→En→Ar direction is for Maltese again, and the worst BLEU

			Direct English		Pivoting through English	
Language Family	**Language X**		**En→X**	**X→En**	**Ar→En→X**	**X→En→Ar**
Finno-Ugric	Hungarian	hu	36.1	48.0	19.1	18.9
	Finnish	fi	38.7	49.5	18.8	19.8
	Estonian	et	38.7	52.2	17.4	20.5
Baltic	Lithuanian	lt	39.2	51.9	19.6	20.4
	Latvian	lv	42.0	54.3	20.7	21.2
Germanic	German	de	46.5	53.5	22.7	21.3
	Danish	da	50.5	57.7	26.2	22.5
	Dutch	nl	52.3	56.8	27.0	22.1
	Swedish	sv	52.2	58.7	24.8	22.7
Greek	Greek	el	49.5	59.5	25.4	23.7
Slavic	Slovak	sk	45.3	61.0	21.9	24.1
	Czech	cs	53.1	58.5	22.4	23.2
	Polish	pl	48.2	61.1	24.3	24.2
	Bulgarian	bg	49.2	61.6	23.7	24.0
	Slovene	sl	51.0	60.9	24.8	24.2
Romance	Romanian	ro	49.2	60.8	25.4	24.0
	Portuguese	pt	55.1	60.6	27.2	23.5
	Italian	it	56.3	61.1	27.8	23.9
	Spanish	es	56.2	60.0	29.8	23.8
	French	fr	62.7	63.7	30.4	25.4
Semitic	Maltese	mt	47.2	72.3	20.5	26.2

Table 2: Pivoting through English and direct English results

score is for Hungarian again. The correlation values between En→X and X→En; and between Ar→En→X and X→En→Ar are not as high: $r = 0.65$ and $r = 0.61$, respectively.

Interestingly, the BLEU scores for En→X are almost double those for Ar→En→X. This is expected but it highlights the need for better MT models for Arabic to Europe's languages.

Correlations Birch et al. (2008) demonstrated that it is possible to predict MT performance using a number of factors: the amount of reordering, the morphological complexity of the target language and the historical relatedness of the two languages. These factors contributed 75% to the variability of the performance of the system.

Our results are consistent with their claims, not only for the direct models which are similar to the models they used but also for those pivoting through English to Arabic. In particular we find the correlation between the word-per-sentence[3] in X to correlate with En→X and Ar→En→X BLEU by $r = 0.82$ and $r = 0.91$, respectively.

However the word-per-sentence does not correlate well when X is the source language: X→En and X→En→Ar by $r = 0.48$ and $r = 0.56$, respectively. Instead we observe that generally the BLEU scores within each family tend to cluster within a small range. Indeed, if we rank the language families in the order shown in Table 2 form 1 to 7, the correlation between this rank and the X→En BLEU and X→En→Ar BLEU are $r = 0.90$ and $r = 0.93$, respectively; while the correlation in the reverse direction does not hold strongly: En→X BLEU and Ar→En→X BLEU correlate with language family rank at $r = 0.75$ and $r = 0.64$, respectively.

[3]The number of words per sentence correlates highly with other measures of morphological complexity like type-to-token ratio ($r = -0.96$). The intuition here is that a language that uses less words to capture the same sentence meaning is more complex morphologically, e.g., while English average sentence length is 27 in our corpus, Arabic's is 22, and Finnish is 18.

6 Conclusions and Future Work

We have presented **Arab-Acquis**, a large professionally translated and publicly available dataset for MT evaluation between 22 European languages and Arabic. We also presented first benchmarking results on translating to and from Arabic for 22 European languages using this dataset.

In the future, we plan to maximize the use of this dataset by using it in improving MT between all of the 22 languages and Arabic in both directions. We also plan to host a shared task on MT evaluation using parts of **Arab-Acquis**.

Acknowledgments

Funding for **Arab-Acquis** was generously provided by a Research Enhancement Fund (REF) grant from New York University Abu Dhabi. We also thank Ahmed El Kholy for assistance in the Arabic detokenization tools.

References

Hassan Al-Haj and Alon Lavie. 2012. The impact of arabic morphological segmentation on broad-coverage english-to-arabic statistical machine translation. *Machine translation*, 26(1-2):3–24.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 745–754, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mohamed Mahdi Boudabous, Nouha Chaâben Kammoun, Nacef Khedher, Lamia Hadrich Belguith, and Fatiha Sadat. 2013. Arabic wordnet semantic relations enrichment through morpho-lexical patterns. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–6, Sharjah, United Arab Emirates. IEEE.

Mauro Cettolo, Nicola Bertoldi, Marcello Federico, and FBK-Fondazione Bruno Kessler. 2011. Bootstrapping Arabic-Italian SMT through comparable texts and pivot translation. In *Proceedings of the 15th Conference of the European Association for Machine Translation (EAMT)*, pages 249–256, Leuven, Belgium.

Ahmed El Kholy and Nizar Habash. 2010. Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC)*, pages 45–51, Valletta, Malta.

Ahmed El Kholy and Nizar Habash. 2012. Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*, 26(1-2):25–45.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 412–418, Sofia, Bulgaria. Association for Computational Linguistics.

European Parliament. 2016. Multilingualism in the european parliament. `http://www.europarl.europa.eu/aboutparliament/en/20150201PVL00013/Multilingualism`. Accessed: 2016-07-15.

Nizar Habash and Jun Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece. Association for Computational Linguistics.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Olivier Hamon and Khalid Choukri. 2011. Evaluation methodology and results for english-to-arabic mt. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 480–487, Xiamen, China.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Philipp Koehn and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of the 12th Machine Translation Summit (MT XII)*, pages 1–8, Ottawa, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 Machine Translation Systems for Europe. In *Proceedings of the 12th Machine Translation Summit (MT XII)*, pages 65–72, Ottawa, Canada.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media, Verlag New York.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, PA, USA. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of LREC*, pages 1094–1101, Reykjavik, Iceland.

Alexandre Rafalovitch, Robert Dale, et al. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the 12th*

Machine Translation Summit (MT XII), volume 12, pages 292–299, Ottawa, Canada.

Reshef Shilon, Nizar Habash, Alon Lavie, and Shuly Wintner. 2012. Machine translation between hebrew and arabic. *Machine translation*, 26(1-2):177–195.

Michel Simard, George F Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, pages 1071–1082, Toronto, Ontario, Canada. IBM Press.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pages 484–491, Rochester, NY, USA. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of NAACL-HLT*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.