# Enhancing Access to Online Education: Quality Machine Translation of MOOC Content

**Valia Kordoni[1], Antal van den Bosch[2], Katia Kermanidis[3], Vilelmini Sosoni[3], Kostadin Cholakov[1], Iris Hendrickx[2], Matthias Huck[4], Andy Way[5]**

[1]Department of English and American Studies, Humboldt University, Berlin, Germany
[2]Centre for Language Studies, Radboud University, Nijmegen, Netherlands
[3]Ionian University, Corfu, Greece
[4]School of Informatics, University of Edinburgh, UK
[5]ADAPT Centre, School of Computing, Dublin City University, Ireland
evangelia.kordoni@anglistik.hu-berlin.de, a.vandenbosch@let.ru.nl, kerman@ionio.gr, vilelmini@hotmail.com,
cholakov@anglistik.hu-berlin.de, i.hendrickx@let.ru.nl, mhuck@inf.ed.ac.uk, away@computing.dcu.ie

**Abstract**

The present work is an overview of the TraMOOC (Translation for Massive Open Online Courses) research and innovation project, a machine translation approach for online educational content. More specifically, videolectures, assignments, and MOOC forum text is automatically translated from English into eleven European and BRIC languages. Unlike previous approaches to machine translation, the output quality in TraMOOC relies on a multimodal evaluation schema that involves crowdsourcing, error type markup, an error taxonomy for translation model comparison, and implicit evaluation via text mining, i.e. entity recognition and its performance comparison between the source and the translated text, and sentiment analysis on the students' forum posts. Finally, the evaluation output will result in more and better quality in-domain parallel data that will be fed back to the translation engine for higher quality output. The translation service will be incorporated into the Iversity MOOC platform and into the VideoLectures.net digital library portal.

**Keywords:** MOOCs, statistical machine translation, crowdsourcing, CrowdFlower, entity recognition, sentiment analysis

## 1. Introduction

Massive Open Online Courses (MOOCs) have been growing in impact and popularity in recent years. According to 2013 statistics[1], more than 200 universities around the globe are involved in their creation, with the participation of more than 1,300 instructors, more than 1,200 courses on offer and around 10 million users being actively enrolled. Apart from their significant contribution to lifelong education, MOOCs are viewed as a tool to help identify and fill the gap that exists in the digital skills of workers across Europe. However, the biggest obstacle standing in the way of further growth in online courses is the language barrier, given that the vast majority of such courses are offered in English.

Although the need for translating MOOC content has been acknowledged by the majority of course providers[2], the solutions provided so far have been fragmentary, human-based, and implemented off-line. TraMOOC [3] constitutes a solution to online course content translation that aims at eleven target languages, is automatic –i.e. it is based on statistical machine translation (SMT) techniques– and is therefore easily extendable to other languages, is adaptable to various types of educational content genre, is independent of course domain, and is designed to produce translations online via its integration in the use-case platforms.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the project. Section 3 describes the collected data, and their domain- and genre-specific idiosyncracies, while Section 4 reports some preliminary translation results. The crowdsourcing tasks involved are presented in Section 5, and the multimodal evaluation schemata and the end product use case plans are described in Sections 6 and 7, respectively.

## 2. Overview of TraMOOC

TraMOOC aims at developing high-quality translation of all types of text genre included in MOOCs (e.g. assignments, tests, presentations, lecture subtitles, forum text) from English into eleven European and BRIC languages, i.e. DE, IT, PT, EL, NL, CS, BG, CR, PL, RU, ZH that constitute strong use-cases, many of which are hard to translate into and have relatively weak MT support. Phrase-based and syntax-based SMT models are developed to address language diversity and support the language-independent nature of the methodology. For high-quality MT and to add value to existing infrastructure, extensive advanced bootstrapping of new resources is performed, while at the same time innovative multi-modal automatic and human evaluation schemata are applied. For human evaluation, an innovative, strict-access control, time- and cost-efficient crowdsourcing set-up is used, while translation experts, domain experts and end users are also involved. Results are combined into a feedback vector and used to refine parallel data and retrain translation models towards a more accurate second-phase translation output. The project results will be showcased and tested on the Iversity MOOC platform and on the VideoLectures.NET digital video lecture library.

[1] https://www.edsurge.com/n/2013-12-22-moocs-in-2013-breaking-down-the-numbers

[2] http://blog.coursera.org/post/50452652317/coursera-partnering-with-top-global-organizations

[3] www.tramooc.eu

The translation engine employed in TraMOOC is Moses[4], the most widely used SMT toolkit available in academia as well as in commercial environments, mainly due to its flexibility, modularity, open-source licence, and competitive translation results.

## 3. Description of the Data

TraMOOC employs a fine-grained definition of the term *educational data* and defines domains in terms of subject areas and text types. Regarding the latter, we make a distinction between formal (lectures subtitles, presentations, assignments and tests) and informal (forum data) domains. Regarding the former, we divide the data into scientific and general-domain. Such distinctions are extremely important, especially with respect to out-of-domain data used to train the SMT system. Although not being strictly educational, parts of the out-of-domain data contain domain-specific terms and phrases which also occur in the TraMOOC domains. For example, out-of-domain parallel corpora derived from OpenOffice and PHP manuals are likely to contain terms which are also frequently used in scientific MOOCs on programming and other IT topics.

To date, there are very few parallel educational corpora for the eleven languages targeted in TraMOOC that are readily available to train SMT systems. One of the significant scientific contributions of TraMOOC is to produce parallel corpora from the educational domain and make them publicly available after the end of the project.

The lack of in-domain parallel data is a real challenge. SMT systems heavily rely on the domains of the parallel data used for training. TraMOOC addresses this issue in two ways: (i) crawling multilingual educational web resources in order to seed the building of parallel corpora, and (ii) building parallel corpora via crowdsourced translation of educational texts. Additionally, we are also exploring the possibility of using parallel corpora from other, closely related languages. For example, we are currently experimenting with adding Serbian parallel data to the already available Croatian parallel corpora. In this way, we can increase the amount of parallel data available for low-resourced languages such as Croatian.

Translated subtitles and other course materials from the Coursera education platform[5] have been one of the major sources for the crawling of parallel educational data[6]. The data compiled so far contains translated materials from over 250 courses offered by Coursera. We have managed to extract translations for all eleven languages targeted in TraMOOC, but the size of the available parallel data varies strongly for each language pair. While for German and Italian there are over 2 million words of parallel data, for Czech and Greek there are only about 200,000 words available.

The translations are produced by course participants who usually translate into their native languages. Translations are done via the Transifex platform[7]. We have developed Python scripts which download the source texts (STs) and the target texts (TTs), i.e. the translations, from Transifex automatically. Since every translation contains a language code, we can easily extract only translations to one of the eleven target languages. Sometimes, there are multiple translations available for a single English segment. In most such cases, Coursera users have voted on the quality of the various translations available and we extract the translation with the highest number of votes.

Regarding the quality of the translations one should keep in mind that this is basically a crowdsourced corpus. Apart from users voting on the quality of the translation, there is hardly any other mechanism for quality assurance. Therefore, we have implemented some basic heuristics for pre-processing and quality checks. For example, we filter out very short segments such as music, silence, applause, etc. Further, we check the length of the source and translated segments. Parallel segments with very large differences in their length are considered dubious and are removed from the corpus. These are mostly segments which for some reason were translated in a wrong language. It is worth noting that Transifex is primarily used for translating subtitles. We found out that the majority of segments represent real, well-formed sentences but the sentences are usually short. Sentence segmentation is therefore generally good, although there are some segments which do not represent well-formed sentences.

Furthermore, there are ongoing experiments with web resources of the EU which can be exploited for all EU languages in the project, e.g. the EU Teacher's Corner[8]. It includes e-books and other electronic educational materials available in many of the 24 official EU languages. All materials aim at educating students from different age groups. The size of the corpus varies for each language because not all materials are available in the same set of languages.

We have also obtained the QCRI Educational Domain Corpus created by the Qatar Computation Research Institute (Abdelali et al., 2014). The corpus consists mostly of educational lectures from the Khan Academy and Udacity educational organisations, but there is also some content from selected Coursera courses. The lectures have been collaboratively transcribed and translated with the AMARA web-based platform[9]. Therefore, this is also a crowdsourced corpus. The data have been cleaned from incomplete subtitles, as well as subtitles which were in a wrong language. Other than that, no further steps for quality assurance have been taken. The corpus contains 20 languages, including 9 TraMOOC languages. There is no parallel data for Croatian and Greek in the corpus.

Last but not least, we also make use of in-domain parallel

[4] http://www.statmt.org/moses/

[5] https://www.coursera.org/

[6] Coursera has provided its consent and has given the TraMOOC consortium access to this data.

[7] https://www.transifex.com

[8] http://europa.eu/teachers-corner/recommended-material/index_en.htm

[9] https://amara.org

data available only for some of the eleven target languages. For example, for German we have obtained parallel data produced within the EU-funded transLectures project (FP7 ICT project #287755)[10]. The data includes transcripts of selected courses available from VideoLectures.NET[11] which were translated by professional translators. The lectures include topics ranging from Computer Science, Technology, Mathematics, Physics, Chemistry and Biology to Business, Social Science and Arts. Although the size of the data is not that large (around 300,000 words), such high-quality parallel data can be very useful for the tuning of the MT models.

Table 1 provides an overview of the size of parallel in-domain data collected so far for each of the eleven TraMOOC languages. The size is given in millions of English words.

Creating parallel corpora via crowdsourcing is another way for obtaining in-domain data which we are pursuing in TraMOOC. We aim at annotating 1 million words per language pair. Due to the use of filtering techniques, like the selection of the best choice among redundant translations, or the automatic detection of errors, in order to ensure the quality of the crowdsourced data, the size of the usable in-domain parallel corpora is expected to be between 800,000 and 850,000 words per language pair. The texts will be carefully selected from subtitles of MOOC courses, course assignments, slides, and other course materials. The forum data of TraMOOC's industrial partner, Iversity, are also included since student forums will also be automatically translated for the purposes of implicit translation evaluation.

| **Language pair** | **Size (million words)** |
|---|---|
| EN-DE | 2.7 |
| EN-BG | 1.5 |
| EN-PT | 4.8 |
| EN-EL | 2.4 |
| EN-NL | 1.3 |
| EN-CZ | 1.5 |
| EN-RU | 1.4 |
| EN-CR | 0.2 |
| EN-PL | 1.7 |
| EN-IT | 2.3 |
| EN-ZH | 8.7 |

Table 1: Size of parallel data for all language pairs

## 4. Initial Translation Results

For the initial TraMOOC prototypes we focused on three language pairs: EN-IT, EN-PT and EN-EL. Phrase-based models were trained on a large amount of parallel and monolingual data, including TED (Cettolo et al., 2012), Europarl (Koehn, 2005), JRC-Acquis (Steinberger et al., 2006), various OPUS corpora (Tiedemann, 2012), WMT News Commentary and Common Crawl[12], and SETimes (Tyers and Alperen, 2010). These were supplemented with monolingual Wikipedia corpora[13] for all three target languages.

The phrase-based models include many features which make them strong baselines. These models include standard features plus a hierarchical lexicalised reordering model (Galley & Manning, 2008), a 5-gram operation sequence model (Durrani et al., 2013), binary features indicating absolute occurrence count classes of phrase pairs, sparse phrase length features, and sparse lexical features for the top-200 words. The models were optimised to maximise BLEU (Papineni et al., 2002) with batch MIRA (Cherry & Foster, 2012) on 1000-best lists.

In Table 1 we compare the BLEU score performance of the systems on the TraMOOC test sets, when tuned on a mixed domain tuning set, or with the TraMOOC tuning set. The mixed tuning set includes tuning sets from TED, Europarl, and News to result in the highest possible general performance system (Huck et al., 2015). As expected, however, it is outperformed by using the domain-specific test set.

| **System** | | **TraMOOC Dev** | **TraMOOC Test** |
|---|---|---|---|
| EN-EL | Tuned Mixed | 25.5 | 28.0 |
| | Tuned TraMOOC Dev | 27.9 | 28.5 |
| EN-PT | Tuned Mixed | 34.1 | 27.9 |
| | Tuned TraMOOC Dev | 36.5 | 29.1 |
| EN-IT | Tuned Mixed | 34.1 | 32.6 |
| | Tuned TraMOOC Dev | 35.9 | 33.0 |

Table 2: BLEU scores for the initial translation prototypes

## 5. Crowdsourcing

Crowdsourcing has been employed extensively for the implementation of human intelligence natural language processing (NLP) tasks in recent years (Callison, 2009; Zaidan & Burch, 2011; Zbib et al., 2012; Ambati, 2012; Finin et al., 2010; Hsueh et al., 2009).

TraMOOC involves crowdsourcing for realizing sub-goals that require human intervention in order to meet its high-quality output standards against upcoming challenges, including the large number of targeted languages, the fragmentary or weak SMT infrastructure support for the majority of the languages, and the multiple domains and text genres involved.

The CrowdFlower[14] platform was chosen for the implementation of the crowdsourcing activities because of (a) its configurability, (b) its robust infrastructure, (c) its densely populated crowd channels and the evaluation and ranking process they undergo, (d) its convenient payment options, and (e) its high reception and popularity level in the microtasking field.

The targeted crowds consist of (a) translation *experts*, (b) an *internal* group of workers with a known background in linguistics and/or translation, and (c) a group of *external*

[10] https://www.translectures.eu/web/project-summary/

[11] http://videolectures.net/

[12] http://www.statmt.org/wmt15/translation-task.html

[13] http://dumps.wikimedia.org (April 2015)

[14] www.crowdflower.com

contributors from the platform's crowd channels. For the latter crowd category, apart from the standard channel evaluation processes applied by the platform to isolate spammers and contributors with poor language skills, further quality assurance measures are taken like

- access control using quiz data that are far from straightforward to address and
- the assignment of each row (for a percentage of the total data rows) to more than one contributors (redundancy).

A separate crowdsourcing task is set up for every language pair and for every NLP activity type. Approximately a total of 2.2 M rows (segments) will be processed. The cost of each activity type varies, depending on its complexity, and is in alignment to the costs reported for similar tasks in the literature.

A microtask in CrowdFlower requires the configuration of several parameter settings that pertain to the number of rows to be tackled in one page, the accepted error rate per page, the maximum number of judgments per contributor, etc. To optimize the configuration a series of trial set-ups have been run before the main tasks, where the participants' comments were recorded and taken into account.

TraMOOC focuses on three types of NLP activities, namely human translation, evaluation of MT output, and text annotation.

### 5.1 Translation

Human translation focuses on the development of in-domain (educational) and in-genre parallel data for training the translation engine, in particular for language pairs that are not adequately equipped with parallel data. This task will be available to internal and external contributors. Each contributor has to translate a set of ten segments in order to complete, submit and get paid for a job. A maximum number of 600 segments have been assigned per contributor. The goal for this task is for the number of segments to be translated per language to exceed 100,000. The cost per segment has been set at 0.04€.

### 5.2 Evaluation

The evaluation task includes several distinct sub-activities, which will form four different crowdsourcing tasks either independently or combined:

1. Likert scale adequacy/fluency marking and post-editing. This task will be opened to internal and external contributors, and will involve approximately 75,000 segments per language pair. The cost per segment has been set at 0.02€.
2. This task includes Task 1 plus error type mark up. Error types will include inflectional morphology, word order, omission, addition and mistranslation. This task will target approximately 15,000 segments per language, and will be mainly carried out by internal contributors and experts. The segment cost has been set at 0.05€.
3. Error taxonomy-based evaluation for translation model comparison for two language pairs, EN-DE and EN-CZ, for 1,000 segments per pair. This task will be mainly carried out by experts. Segment cost has been set at 0.05€.
4. Ranking multiple translations of a given segment. Redundant translations provided in the translation task will be used in this evaluation task. External and internal contributors will be asked to rank the provided translation in decreasing quality. This task will target around 8,000 segments per language pair. Segment cost has been set at 0.02€.

Experimentation with various crowd types and comparative testing between different task complexity levels aims at investigating the usability, the usefulness and the efficiency of sophisticated human evaluation.

### 5.3 Annotation

Text annotation involves two different crowdsourcing tasks: entity annotation and sentiment annotation. The former will be applied to 1,000 segments per language pair, for all eleven pairs. Each segment will be annotated three times by three distinct contributors. The annotation process includes the markup of a potential single- or multi-word entity in the source segment, the linking with its Wikipedia URL (if available), and then the parallel process in the target segment. The segment cost has been set at 0.05€.

Sentiment annotation will be applied to English segments only, taken from the MOOC forum students' text. Contributors will identify whether a given segment contains a positive, neutral or negative opinion regarding the machine-translated course content, produced by the TraMOOC translation engines. The cost for each segment annotation is 0.01€.

The aforementioned annotations are used for training and/or testing the entity recognition and the sentiment analysis tools; the output of these tools facilitates the implicit evaluation setup described in detail in the next section.

## 6. Evaluation Schema

### 6.1 Explicit Evaluation

Explicit evaluation involves automatic and human evaluation of the translation output. In particular, n-gram similarity-based metrics, e.g. BLEU or NIST (Papineni et al., 2002), or word-editing based metrics, e.g. TER[15], are used for estimating the accuracy of the translated text. Diagnostic evaluation is performed focused on specific linguistic phenomena and error types. Comparative analysis of the results is performed across translation models, across language models, across languages, and across text types. Human evaluation is performed via crowdsourcing, as described earlier, and involves domain experts, translation experts and non-experts.

[15] http://www.cs.umd.edu/~snover/tercom/

The comparative analysis of translation models will comprise automatic and human evaluation of syntax-based SMT, phrase-based SMT, and Neural MT, in the English-German and English-Czech language pairs. This stage will also provide a valuable linguistic checkpoint for ST issues. Thereafter, human evaluation by experts and non-experts will rate quality, highlight commonly occurring errors in MT output, and provide edited TT segments for retraining the MT engines in order to improve quality and domain specificity.

### 6.2 Implicit Evaluation

Implicit MT evaluation aims to judge the MT quality between the ST and automatically generated TT without using a manually created reference translation. In TraMOOC, topic identification and sentiment analysis will be used for this task. Topic identification is performed via wikification (Mihalcea, 2007). Sentiment analysis extracts the opinion of end-users regarding the TT by applying opinion mining techniques to user contributions posted on the MOOC forum.

For the implicit MT evaluation we focus on topical information elements (named entities, events, specific terms) in source and target documents. Topic identification can be done in several ways such as computing word weights (Wartena et al., 2010), using the document structure to find the main topics (Hearst, 1995), or applying an (un)supervised topic modeling technique such as LDA (Blei et al., 2003).

In TraMOOC we make use of the fact that most Wikipedia pages have translations in many other languages, and use wikification for implicit MT evaluation. Using name translation as a measure for overall MT quality has been suggested before and has been shown to correlate well with human MT judgments (Hirschman et al., 2000). With wikification we aim to generalize this technique. Such a wikification system detects both named entities and terms (topics) in a document and links them to their corresponding Wikipedia page. We apply a Wikifier (Ratinov & Roth, 2011) to find and link the topics in the English source data to their relevant Wikipedia pages. Next, we use the alignment between the source and target sentence to get the corresponding translation of the topics in the TT. We check whether this translated topic corresponds to the same Wikipedia page in the ST. When such a match is found, we count this as a correct topic translation, and when no matching page is found we count it as an error. The transformation of the Wikification results into a reliable implicit evaluation MT score is a crucial research question that we will pursue in the course of the project.

We create a reference set for the tuning and testing of the Wikifier and the development of the implicit score metric. For this reference set we collected 1000 sentences from MOOC courses in the eleven languages. These sentences are manually annotated with Wikipedia links via the Crowdsourcing platform. Each sentence is annotated by three different annotators and only annotations supported by at last two annotators are kept in the final reference set. This set will also give us an indication of the limits of the wikification method, such as its dependency on coverage of topics per language. Greek, for example, only has around 115,000[16] Wikipedia pages available and many detected topics in the English will not have corresponding Greek Wikipedia pages. Therefore we expect a lower coverage of the implicit evaluation method for low-resource languages as can be illustrated in the following example.

In Example 1 we show a sentence taken from the Iversity MOOC course on Critical Thinking. Examples 2 and 3 show the automatic translation of this sentence in Portuguese and Greek produced by the prototype-1 TraMOOC MT system. In both cases the translation of the name is only partly correct. In Portuguese the correct translation should have been ‘Trilema de Münchhausen’ or the synonym ‘Trilema de Agripa’ that both point to the same existing Wikipedia page. Due to the incorrect translation, the implicit MT evaluation will count this as a translation error. For Greek, no equivalent Wikipedia page exists and the translation quality of ‘Ο Αγρίππας του Μυγχάουσεν’ cannot be verified.

*(1) Agrippa's Trilemma states that there are three options if we try to prove any truth .*

*(2) PT: Agrippa Munchausen afirma que existem três opções se tentarmos provar qualquer verdade.*

*(3) EL: Ο Αγρίππας του Μυγχάουσεν αναφέρει ότι υπάρχουν τρεις επιλογές αν προσπαθήσουμε να αποδείξει κάποια αλήθεια.*

## 7. Use Cases

The technology developed in the TraMOOC project is applied to two different use-cases: the European MOOC platform, Iversity, and the VideoLectures.NET digital video lecture library.

Iversity is a Berlin-based MOOC provider which launched its first MOOCs in 2013 and has grown quickly, recently reaching a cumulative 500,000 users with over 700,000 course enrollments. There are now ~50 courses from a few dozen European universities, with most users in Europe, but also from all parts of the world[17]. Courses cover areas ranging from Design, Engineering, and Computer Science to Education, Philosophy, and Life Sciences. The language of the vast majority of courses provided is English, and courses are mostly held via video lectures, though some have additional textual material, such as slides. All courses are accompanied by a forum platform that allows students and teachers to communicate. The translation prototypes generated in TraMOOC will be integrated into the Iversity platform according to the end-user requirements.

VideoLectures.NET, administered by the Knowledge 4 All Foundation Ltd. and run by the dedicated Center for Transfer in Information Technologies at the Josef Stefan

[16] verified at 23-02-2016

[17] https://www.class-central.com/report/iversity-european-moocs/

Institute (JSI) in Ljubljana, was founded in 2001. It functions as a free, online video library, established with the aim of promoting access to academic lectures given by distinguished scholars, scientists, researchers and academics from many scientific fields, at conferences, summer schools, workshops, and university classrooms. Lecture subtitles translated via the TraMOOC translation prototypes will be accessible through the video library.

## 8. Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 644333, as well as from the ADAPT Centre, Dublin.

## 9. References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA Corpus: Building parallel language resources for the educational domain. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC'14). Reykjavik, Iceland, pp. 1856-1862.

Ambati, V. (2012). *Active learning and crowd-sourcing for machine translation*. PhD Thesis. Carnegie Mellon University. ISBN: 978-1-267-58215-7.

Blei D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3), pp. 993–1022.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*. Association for Computational Linguistics, pp. 286-295.

Cettolo, M., Girardi, C. and Federico, M. (2012). WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Annual Conference of the European Association for Machine Translation* (EAMT). Trento, Italy, pp. 261–268.

Cherry, C. Foster, G. (2012). Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics* (HLT-NAACL). Montréal, Canada, pp. 427–436.

Durrani, N., Fraser, A. and Schmid, H. (2013). Model With Minimal Translation Units, But Decode With Phrases. In *Proceedings of the Human Language Technology Conference North American Chapter of the Association for Computational Linguistics* (HLT-NAACL), Atlanta, GA, USA, pp. 1–11.

Finin, T. Murnane, W., Karandikar, A., Keller, N. and Martineau, J. (2010). Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, pp. 80-88.

Galley, M., Manning, C. D. (2008). A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing* (EMNLP). Honolulu, USA, pp. 847–855.

Hearst, M. A. (1995). Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Denver, CO, USA: ACM Press/Addison-Wesley Publishing Co, pp. 59–66.

Hirschman, L., Reeder, F., Burger, J. D. and Miller, K. (2000). Name translation as a machine translation evaluation task. In *Proceedings of the Workshop Evaluation of Machine Translation*, LREC. Athens, Greece, pp. 21-28.

Hsueh, P., Melville, P. and Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics, pp. 27-35.

Huck, M., Birch, A. and Haddow, B. (2015). Mixed-Domain vs. Multi-Domain Statistical Machine Translation. In *Proceedings of MT Summit XV, vol.1: MT Researchers' Track*. Miami, FL, USA, pp. 240-255.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit X*. Phuket, Thailand, pp. 79-86.

Mihalcea, R. and Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, CIKM '07. New York, NY, USA, pp. 233–242.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (ACL). Philadelphia, PA, USA, pp. 311–318.

Ratinov, L., Roth, D., Downey, D. and Anderson, M. (2011). Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume* 1. Portland, Oregon, USA, pp. 1375–1384.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. and Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC). Genoa, Italy, pp. 2142–2147.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC). Istanbul, Turkey, pp. 2214–2218.

Tyers, F. M. and Alperen, M. S. (2010). South-East European Times: A parallel corpus of Balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*. Malta, pp. 49–53.

Wartena, C., Brussee, R. and Slakhorst, W. (2010). Keyword extraction using word co-occurrence. In *Proceedings of the 23rd International Workshop on Database and Expert Systems Applications*. Bilbao, Spain, pp. 54–58. doi:10.1109/DEXA.2010.32

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 1220-1229.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F. and Callison-Burch, C. (2012). Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 49-59.