MEDLINE as a parallel corpus: a survey to gain insight on French-, Spanish- and Portuguese-speaking authors' abstract writing practice

Aurélie Névéol, Antonio Jimeno Yepes, Mariana L Neves

MEDLINE as a parallel corpus: a survey to gain insight on French-, Spanish- and Portuguese-speaking authors' abstract writing practice

Aurélie Névéol[†], **Antonio Jimeno Yepes**[*], **Mariana Neves**[⋆]
[†] Université Paris Saclay, CNRS, LIMSI, France
[*] IBM Research Australia
[⋆] German Federal Institute for Risk Assessment (BfR), Germany
aurelie.neveol@limsi.fr, antonio.jimeno@au1.ibm.com
mariana.lara-neves@bfr.bund.de

Abstract

Background: Parallel corpora are used to train and evaluate machine translation systems. To alleviate the cost of producing parallel resources for evaluation campaigns, existing corpora are leveraged. However, little information may be available about the methods used for producing the corpus, including translation direction. Objective: To gain insight on MEDLINE parallel corpus used in the biomedical task at the Workshop on Machine Translation in 2019 (WMT 2019). Material and Methods: Contact information for the authors of MEDLINE articles included in the ENES, ENFR and ENPT WMT 2019 test sets was obtained from PubMed and publisher websites. The authors were asked about their abstract writing practices in a survey. Results: The response rate was above 20%. Authors reported that they are mainly native speakers of languages other than English. Although manual translation, sometimes via professional translation services, was commonly used for abstract translation, authors of articles in the esen and pten sets also relied on post-edited machine translation. Discussion: This study provides a characterization of MEDLINE authors' language skills and abstract writing practices. Conclusion: The information collected in this study will be used to inform test set design for the next WMT biomedical task.

Keywords: Parallel corpus, biomedical NLP, translation quality

1. Introduction

In the biomedical domain, information is typically provided in English through scientific publications and patient oriented fact sheets prepared by government institutions. Due to the large number of publications that is becoming available, machine translation methods support making this information available in languages other than English. Translation from a variety of languages into English is also important in biomedicine. For instance, it can support researchers working on clinical reports, which are usually available in the national or regional language of the territory where the patient is treated. Given the lack of natural language processing (NLP) tools available for languages other than English, translation can be envisaged as a preprocessing step (Campos et al., 2017).

However, machine translation methods rely on the availability of large parallel corpus for training, tuning and evaluation (Koehn, 2009). While characteristics of language pairs, such as morphological complexity of the target language and relatedness of the two languages, are major indicators of success for translation systems (Birch et al., 2008), training corpus size is a bottleneck, especially for modern neural translation systems (Koehn and Knowles, 2017). The use of triangulation has been explored to compensate for the lack of parallel data in a specific language pair (Gispert and Mariño, 2006). Crowdsourcing has also been used for producing parallel corpus in rare language pairs (Zbib et al., 2012). In spite of these efforts, parallel corpus are still needed, especially in specialized domains.

A parallel corpus of scientific abstracts was collected from MEDLINE for use in the biomedical task offered at the Workshop on Machine Translation (WMT) in 2018 and 2019 (Neves et al., 2018; Bawden et al., 2019). The abstracts in the corpus were produced by the authors of the articles indexed in the MEDLINE database. Little is known about the writing practices used by the authors to create these texts. We have made the hypothesis that authors write their abstract in one language, and then translate it into one or more languages according to the journal requirements. Assuming this hypothesis is correct, there is no information on the language used to write the original abstracts. However, Machine Translation (MT) research has shown that features such as the translation direction have an impact on the performance of translation models trained or tuned on parallel corpus (Kurokawa et al., 2009; Lembersky et al., 2013; Stymne, 2017). Recent investigation of the test sets used in the WMT news translation tasks from 2016 to 2018 found that the use of mixed translation directions in test sets inflates the human scores for translation systems to the level that system rankings can be impacted (Zhang and Toral, 2019). While evaluations carried out on tests sets including documents with mixed translation directions can be re-interpreted in light of the translation direction information, it is now recommended to move away from the use of such test sets (Graham et al., 2019).

Furthermore, direct analysis of the MEDLINE datasets shows that alignment and translation quality are uneven (Névéol et al., 2018). It has been hypothesized that the authors may have no translation training and sometimes lacking competence in some of the languages that they are required to use. Other conjectures include the possibility that authors may write the abstracts in different languages independently to leverage the language competence of different authors in a group, which would put an emphasis on maximizing language correctness and content accuracy rather than translation quality.

This study aims to characterize MEDLINE authors' language competence and writing practice of abstracts. Findings are expected to shed some light on the results of past biomedical tasks results and to inform the use of MEDLINE datasets for future editions of the task.

2. Material and Methods

2.1. Study corpus.

The corpus used in this study comprises the test sets distributed in the WMT 2019 biomedical task (Bawden et al., 2019) for the language pairs English/Spanish (EN/ES), English/French (EN/FR) and English/Portuguese (EN/PT). Each dataset is composed of a total of 100 abstract pairs. Table 1 presents excerpts from the corpus illustrating the diversity of the contents in terms translation styles: litteral translations such as (1), creative translations such as (2), and non idiomatic translations such as (3)[1]. We selected the test set containing the most recent documents used in the WMT biomedical task (articles in the datasets were published in 2019), to maximize the chance that we could reach the authors and that they would adequately remember the writing details pertinent to their articles.

2.2. Collecting author contact information.

MEDLINE citations were used as the primary source for author contact information. When it was not found, the publisher's website was scanned when a DOI was available. When the authors' contact information was still not found, the first and/or last authors' other publications in MEDLINE were used to retrieve an email address either from MEDLINE or the publisher using the same methods as before.

2.3. Survey of abstract writing practice.

A survey was developped and implemented in LimeSurvey[2]. In order to maximize participation, we limited the number of survey questions (four questions, plus two optional comment fields). The original survey questions were written in English, the common language between the study authors, and then translated into French, Portuguese and Spanish. Similarly, we created versions of the invitation message in multiple languages. The authors were contacted by email using the collected contact information. The message was sent in the two languages pertaining to the the language pair of the article. An initial invitation message was sent, as well as a follow-up message around one week later. We used the information from MEDLINE citations (content of the `TT` field) to infer the primary language of the article, so that the invitation emails could be customized to present the message in the primary language first. We hoped that this would help to engage authors and yield a higher response rate. A copy of all survey questions and invitation messages is available from the WMT 2019 biomedical task shared folder [3]. Figure 2 presents the English version of the survey.

[1] A better translation could be: *The LT was **located in the glottis** in all cases (9/9).*

[2] `https://www.limesurvey.org/`

[3] `https://drive.google.com/drive/u/1/folders/1tqaAVB-kk2HTZX9YUCQRmtsgAqW3agfR`

3. Results

3.1. Email collection for the French, Portuguese and Spanish WMT 2019 biomedical task test sets

Email collection was time-consuming, especially for the ENFR set for which no contact information was directly available in the original MEDLINE citations -29 contact emails could be retrieved directly from the original MEDLINE citation for each of the ENPT and ESEN sets. While the majority of the contact emails collected for each test set indicated authors were affiliated with an institution in a country where the Language other than English is spoken (e.g. France, Canada, Switzerland or Belgium for French, Brazil or Portugal for Portuguese, Mexico or Spain for Spanish), a number of the contact emails pointed towards countries where these languages are not official languages (e.g. Germany, Austria, Pakistan, Iran, ...).

We detail the results for each test set in figure 1. Interestingly, the distribution of email domains is different for each language pair. In ENFR, more than half of the email domains are from French speaking countries, while in the other two language pairs, generic email domains are more than half of the total. ENPT is the language pair with smaller proportion of emails from Portuguese speaking countries.

For the ENFR set, contact email was successfully retrieved for 90 PMIDs (out of 100). For an additional PMID, authors could be contacted through a contact form link on the publisher's website. For the ENPT set, contact email was successfully retrieved for all 100 PMIDs. For the ENES set, contact email was successfully retrieved for 90 PMIDs. In some cases, we obtained emails for more than one author.

3.2. Survey responses

Responders and response rate. Table 3 presents the response statistics to the surveys. Most of the responders used the link to the non-English version of the survey (77 % of responders; 68 out of 88 responders in total), which tends to indicate that they are more comfortable with French, Spanish or Portuguese. Overall, the response rate was between 25% and 35%. Some responders did not fill the form completely (18 out of 88 responders in total[4]) but in many cases the portion of the survey affected by the lack of responses was the optional PMID and comment fields. We consider the response level to be reasonable, taking into account that we also received "out-of-office" and "undeliverable mail" responses. (specifically, 5 each for the ENFR set, 5 overall for the ENPT set and 4 undeliverable mails for the ENES set). In addition, authors routinely receive spam email related to their publications and our survey invitation could have been dismissed as such.

Response time. Table 4 presents the average response time to each of the surveys. Overall, response time was well under four minutes, which achieves our goal to limit the demand on responder time.

[4] incomplete responses explain minor inconsistencies in the total number of responses reported in Figures 2 - 5

#	Source	Translation
(1)	Se identificaron múltiples reportes de caso de eventos adversos aislados.	Several cases of isolated adverse events were identified.
(2)	There is no workforce development without workforce intelligence.	On ne saurait parler de développement des personnels si l'on ne dispose pas d'informations à ce sujet.
(3)	La TL était de siè toge glottique dans tous les cas (9/9).	* The LT was of glottic seat in all the cases (9/9).

Table 1: Excerpts from the ENFR and ENES corpus.

1. Who wrote the English abstract for this paper? - One author, who is a native English speaker - One author, who is NOT a native English speaker - A group of authors, including a native English speaker - A group of authors, NOT including a native English speaker - I do not recall
2. Who wrote the Language-other-than-English abstract for this paper? - One author, who is a native Language-other-than-English speaker - One author, who is NOT a native Language other than English speaker - A group of authors, including a native Language-other-than-English speaker - A group of authors, NOT including a native Language-other-than-English speaker - I do not recall
3. What order were the abstracts written in? - English first - Language-other-than-English first - Abstracts were written independently (e.g. by different authors) - I do not recall
4. What was the writing method for the abstracts? - One abstract was written and then translated manually in the other language - One abstract was written and then translated automatically (e.g. using Google Translate) into the other language, without revision - One abstract was written and then translated automatically (e.g. using Google Translate) into the other language, then revised - Abstracts were written independently - I do not recall
Do you have questions or comments that you would like to share regarding your scientific writing practice? *free text field*
Please enter the PMID of the article covered by your answers. Please note that entering this information is optional is will waive the anonimity of your participation to the survey. *Free text field*

Table 2: Survey questions in English. A version of the suvey in Portuguese, French, and Spanish was also provided.

Set	Contacts	All responses	EN responses	Response rate
ESEN	90	31	6	34 %
FREN	91	24	9	26 %
PTEN	100	33	5	33 %

Table 3: Responses to the surveys.

Set	ES/FR/PT survey	EN survey
ESEN	2 min and 48 sec	3 min and 25 sec
FREN	2 min and 1 sec	2 min and 19 sec
PTEN	1 min and 53 sec	1 min and 30 sec

Table 4: Response time statistics in minutes (min) and seconds (sec) (average time to completion).

Authors' language competence. Figure 2 presents the responses regarding the author of the abstract in the language other than English. According to the responders, writing the French, Spanish or Portuguese abstract commonly involved a native speaker of the language (overall, 60 out of 75 responders i.e. 80%) and rarely had no native speaker involved (overall, 5 cases out of 75, i.e. less than 7%). Some of the "I do not recall" responses can be linked

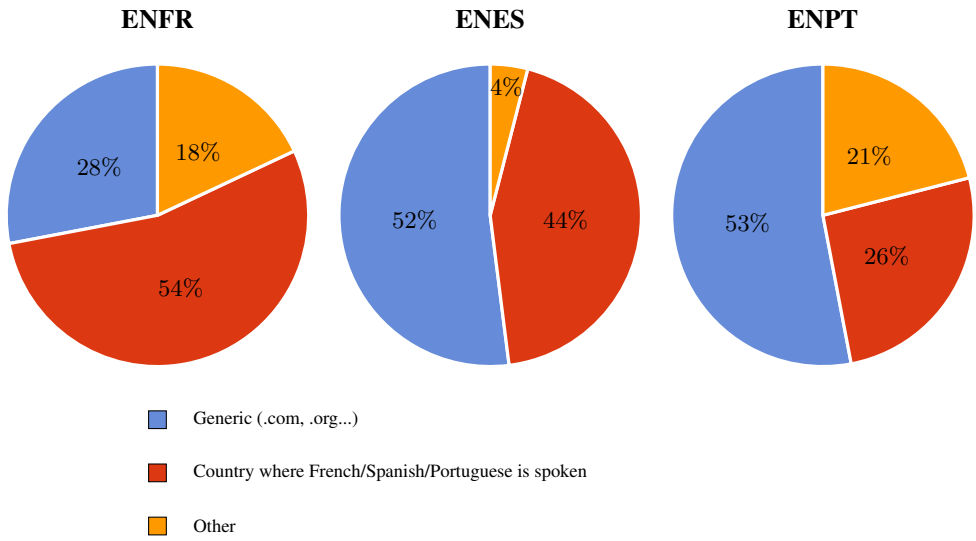


Figure 1: Distribution of authors' email domains

to the comments that a translation of the English abstract was carried out by a third party independent of the authors (e.g. journal of institution translation provider).

Figure 3 presents the responses regarding the author of the English abstract. According to the responders, writing the English abstract commonly did NOT involve a native speaker of English (59 cases out of 76, i.e. almost 78%) and sometimes involved a native speaker of English (11 cases out of 76, i.e. 14%).

Authors' abstract writing practice. Figure 4 presents the responses regarding the order of abstract writing. As could be expected, the responders to the French, Spanish and Portuguese surveys most commonly wrote the language other than English abstract first (41 out of 56 cases, i.e. 73%) while the FREN responders to the English survey most commonly wrote the English abstract first (6 out of 9, almost 70%), results are not so clear for the ESEN and PTEN responders. However, we can note that the low response rate may limit the interpretation of the data.

Figure 5 presents the responses regarding the method used to obtain an abstract in a second language. Manual translation was the most common method (overall, 37 cases out of 72, i.e. 51%). We can also note that automatic translation was not uncommon (21% of responses, including one case where machine translation was not post-edited). Authors in the ESEN and PTEN sets accounted for the large majority of machine translation use (15 out of 16 cases, i.e. 94%). The "I do not recall" responses in the FREN set can be linked to the comments that a translation of the English abstract into French was carried out by a third party independent of the authors (e.g. journal of institution translation provider).

PMIDs and comments. The survey offered the authors the option to provide the PMID of the abstract they were reporting on as well as free text comments regarding any aspect of their abstract writing practice. A total of 26 PMIDs were supplied by respondents accross all datasets. In addition, 12 comments where collected accross surveys, and sometimes expressed the same views regardless of the dataset authors.

The comments essentially addressed two topics. First, they reported on the access of authors to translation services. As summarized below, authors' access to translation services through their institution or through the journal varied from no access (but access was desired) to full access to translation services that they could intereact with:

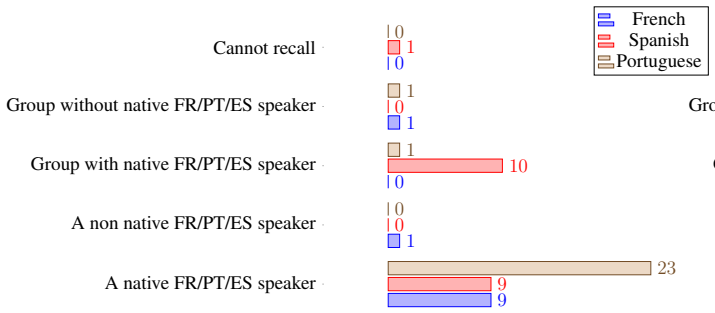
- Authors should have access to professional translation services for abstract translation, ideally provided by the journal requesting the translation as the authors are not qualified.
- Abstracts are written in the author's language of choice (English or French) and the author's [Canadian] institution arranges professional translation into the other language.
- The journal supplied a translation of an abstract originally writen in English by the authors, and the authors were given the opportunity to review the translation or provide the translator with feedback to ensure quality translation.

Second, they reported on the language resources used by the authors to write an abstract in a foreign language:

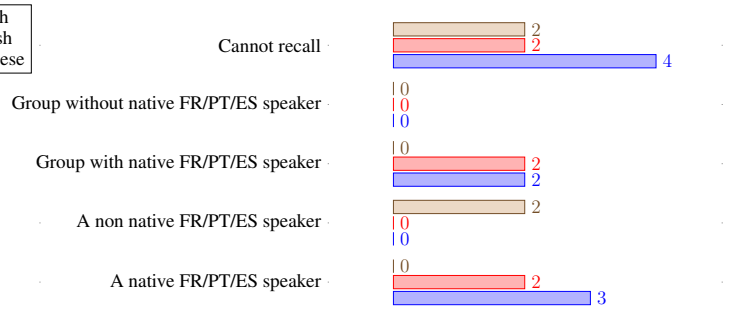
- The group of authors included a native speaker of both languages involved so the translation was not a problem.
- One author reported using Grammarly for writing the English abstract.

4. Discussion

Limitations. A limitation of this work is the small scale of MEDLINE parallel corpus that was addressed (only 300 articles alltogether), and the relatively low response rate (about 30%). Nonetheless, the data collected suggests interesting trends and characteristics of abstract writing practices.

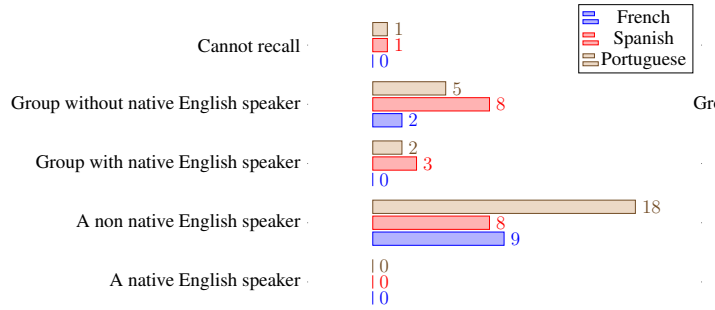


(a)

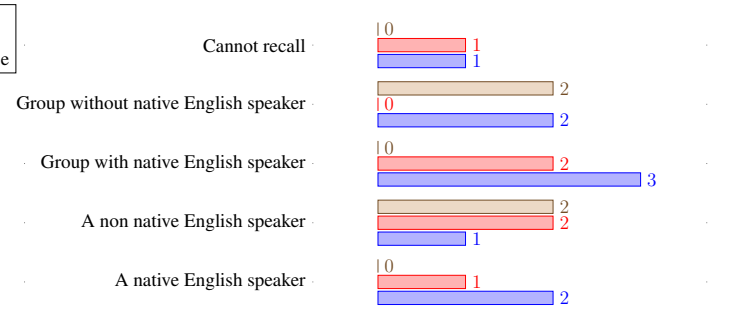


(b)

Figure 2: Author of the French (FR)/Portuguese (PT)/Spanish (ES) abstract. (a) presents the response from the FR/PT/ES survey and (b) from the English survey

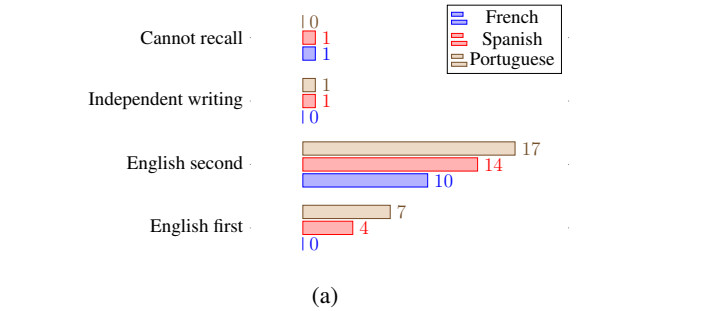


(a)

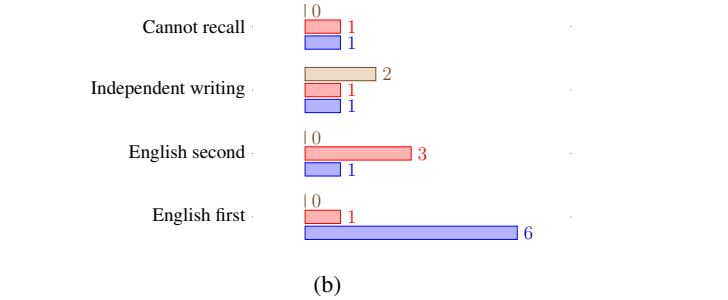


(b)

Figure 3: Author of the English abstract. (a) presents the response from the French (FR)/Portuguese (PT)/Spanish (ES) survey and (b) from the English survey

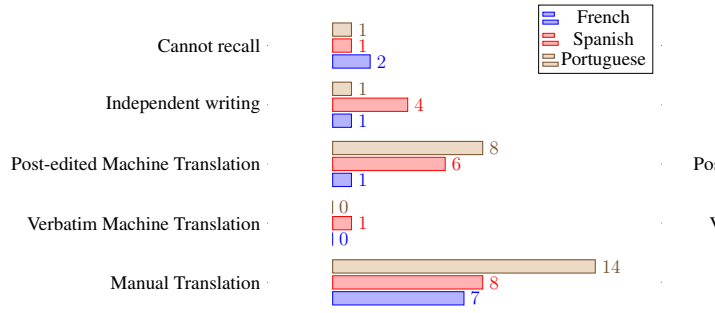


(a)

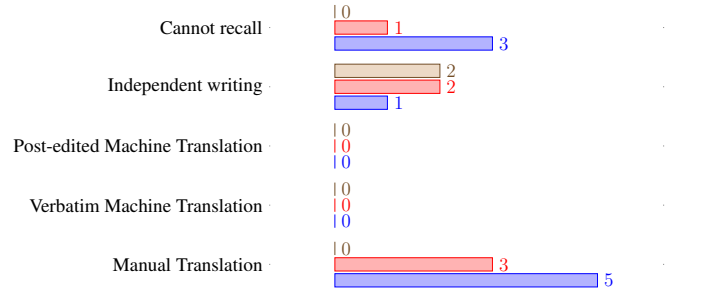


(b)

Figure 4: Order of abstract writing. (a) presents the response from the French/Portuguese/Spanish survey and (b) from the English survey



(a)



(b)

Figure 5: Method for abstract writing. (a) presents the response from the French/Portuguese/Spanish survey and (b) from the English survey

Findings. Figures 2 - 5 show that most of the abstracts have been manually translated. If a machine translation method was used, it would typically be post-edited, which reflects the new standard in the translation industry. Authors typically write abstracts in their native language first, which seems to be mostly French, Spanish or Portuguese for our survey responders and then translate it into English. Our study suggests that Spanish- and Portuguese- speaking authors more readily use technology (use of machine translation, availability of email contact) than French speaking authors.

English abstracts were commonly prepared by a non-native speaker or by a group not including a native English speaker. This might affect the quality of the English translation. Similarly the native language abstract would be written by a native language speaker. The quality of the abstract written by a native speaker will be less problematic than the translation into English by non-English native speakers.

Usability of MEDLINE parallel corpus at WMT. The fact that authors in the ESEN and PTEN sets use post-edited machine translation to produce abstract may contribute to high BLEU scores observed for the language pairs ES/EN and PT/EN. The number of responses providing the specific PMID of the article concerned by survey answers was too low to sufficiently enrich the datasets with information on translation direction and translation method. Similarly, direct survey may yield to sparse data for the creation future test sets. However, based on the responses to the survey (in particular, responses from the authors of the ENFR set) our study suggests that it is possible to infer translation direction for the abstract based on the language used in the original article [5]. This information should be used to adjust the translation direction of future test sets.

5. Conclusion

The information collected in this study will be used to inform test set design for the next WMT biomedical task. In particular, results suggest that the translation direction can be inferred based on the language of the original article, and test sets can be designed accordingly instead of using random split as was done previously.

Acknowledgements

We would like to thank Dr. Rachel Bawden for her insightful comments on the development of survey questions. We are also grateful to the authors who kindly took the time to respond to our survey. This work was partially supported by the Agence Nationale pour la Recherche (French National Research Agency) under grant number ANR-15-CE23-0025-0

References

Bawden, R., Bretonnel Cohen, K., Grozea, C., Jimeno Yepes, A., Kittner, M., Krallinger, M., Mah, N., Neveol, A., Neves, M., Soares, F., Siu, A., Verspoor, K., and Vicente Navarro, M. (2019). Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy, August. Association for Computational Linguistics.

Birch, A., Osborne, M., and Koehn, P. (2008). Predicting success in machine translation. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 745–754. Association for Computational Linguistics.

Campos, L., Pedro, V., and Couto, F. (2017). Impact of translation on named-entity recognition in radiology texts. *Database*, 2017, 08. bax064.

Gispert, A. D. and Mariño, J. B. (2006). Statistical machine translation without parallel corpus: bridging through spanish. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 65–68. European Languages Resources Association (ELRA).

Graham, Y., Haddow, B., and Koehn, P. (2019). Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press.

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *In Proceedings of MT-Summit XII*, pages 81–88.

Lembersky, G., Ordan, N., and Wintner, S. (2013). Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023.

Névéol, A., Jimeno Yepes, A., Neves, M., and Verspoor, K. (2018). Parallel corpora for the biomedical domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Neves, M., Jimeno Yepes, A., Névéol, A., Grozea, C., Siu, A., Kittner, M., and Verspoor, K. (2018). Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels, October. Association for Computational Linguistics.

Stymne, S. (2017). The effect of translationese on tuning for statistical machine translation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 241–246, Gothenburg, Sweden, May. Association for Computational Linguistics.

[5] This may seem trivial, but our hypothesis was that the abstract may have been written independently from the article: the abstract could originally be written in English even though the article itself was in another language.

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*, pages 49–59, Montréal, Canada, June. Association for Computational Linguistics.

Zhang, M. and Toral, A. (2019). The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy, August. Association for Computational Linguistics.