Parallel Corpora for the Biomedical Domain

Aurélie Névéol, Antonio Jimeno Yepes, L Neves, Karin Verspoor

Parallel Corpora for the Biomedical Domain

Aurélie Névéol[†], Antonio Jimeno Yepes[*], Mariana Neves[‡], Karin Verspoor[⋆]
[†] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
[*] IBM Research Australia
[‡] German Federal Institute for Risk Assessment (BfR), Germany
[⋆] University of Melbourne, Australia
aurelie.neveol@limsi.fr, antonio.jimeno@au1.ibm.com
mariana.lara-neves@bfr.bund.de, karin.verspoor@unimelb.edu.au

Abstract

A vast amount of biomedical information is available in the form of scientific literature and government-authored patient information documents. While English is the most widely used language in many of these sources, there is a need to provide access to health information in languages other than English. Parallel corpora can be leveraged to implement cross-lingual information retrieval or machine translation tools. Herein, we review the extent of parallel corpus coverage in the biomedical domain. Specifically, we perform a scoping review of existing resources and we describe the recent development of new datasets for scientific literature (the EDP dataset and an extension of the Scielo corpus) and clinical trials (the ReBEC corpus). These corpora are currently being used in the biomedical task in the Conference on Machine Translation (WMT'16 and WMT'17), which illustrates their potential for improving and evaluating biomedical machine translation systems. Furthermore, we suggest additional applications for multilingual natural language processing using these resources, and plan to extend resource coverage to additional text genres and language pairs.

Keywords: Parallel corpus, biomedical domain, multilingual applications

1. Introduction

Machine translation (MT) is currently being used for a variety of tasks and domains. It is known to play an important role in supporting readers' access to textual documents in a language other than their native language or for communicating in real time. The accuracy of MT systems has improved in recent years thanks to the availability of large collections of parallel and/or comparable corpora. In turn, these resources could be leveraged by deep learning methods, which created a paradigm shift for MT.

MT plays an important role in the health domain. For instance, it has the potential to enable patients to read documents written in a language in which they are not fluent and to hold a conversation with foreign health professionals in case of accidents or health issues in a foreign country. Further, it allows patients to access health information which is only available in a foreign language, for instance, in the case of disease outbreak with origin in other countries (e.g., Zika virus outbreak in Brazil (Bueno, 2017)).

MT can also support researchers to access scientific literature only available in a foreign language, for instance, when working on tropical diseases specific of a region or even when moving to another country for research purposes (Walker, 2016). Finally, MT can also support the biomedical natural language processing (BioNLP) domain when processing documents in languages other than English for which no specific NLP tools are available. This is often the case for clinical discharge reports that are usually only available in the local language. In such cases, researchers could translate the original document into English and rely on state-of-the-art BioNLP tools that are available for English (**?**). The biomedical and health domain is well known for its complex nomenclature, for which specific language resources and tools have been developed, e.g., lemmatizers (Liu et al., 2012). Therefore, specific training and test datasets are also necessary to precisely translate biomedical document across languages. However, despite its importance for the general population and researchers, there are very few parallel and comparable corpora specific for this domain.

In this paper, we present an overview of the state-of-the-art on parallel and comparable corpora for the biomedical domain. In a scoping review of existing resources, we characterize the resources available by language pairs and document type and provide pointers to more in-depth descriptions of the resources. Additionally, we present the parallel corpora that we assembled and built, such as EDP (French/English), ReBEC (Neves, 2017) (Portuguese/English) and Scielo (Neves et al., 2016) (French/English, Portuguese/English and Spanish/English). For the latter, we provide details on the corpus construction, insights on the data and their utilization for the biomedical task (Bojar et al., 2016; Jimeno Yepes et al., 2017) of the Conference for Machine Translation (WMT). All corpora are available in our repository in GitHub[1].

2. Related Work

One of the first efforts that involved the development of large-scale shareable parallel corpora for the biomedical domain was the OPUS collection that contained medical documents from the European Medicines Agency (EMEA) (Tiedemann, 2012)[2]. A number of biomedical parallel (Widdows et al., 2002; Ozdowska et al., 2005; Deleger et al., 2009) and comparable corpora (Chiao and Zweigenbaum, 2004) have been used for terminology translation only. Similarly, the Mantra project (Kors et al., 2015;

[1] http://github.com/biomedical-translation-corpora/corpora
[2] `http://opus.lingfil.uu.se/EMEA.php`

Hellrich et al., 2014) provided corpora of biomedical articles automatically annotated for named-entity recognition for English, Spanish, French, German and Dutch. The corpora included MEDLINE titles, EMEA documents and patents in the biomedical field. While the goal of this project was to leverage annotation transfer from English to other languages to expand terminology coverage in languages other than English, to our knowledge, the corpus has not been used for machine translation.

After general purpose machine translation systems were found to perform poorly on medical text (Zeng-Treitler et al., 2010), the use of domain-specific data was investigated to improve MT system performance in the biomedical field. MEDLINE titles and terms from the Unified Medical Language System (UMLS) (Lindberg et al., 1993) were investigated first, due to easy availability (Wu et al., 2011; Jimeno Yepes and Névéol, 2013). Abstract sentences were also found useful but difficult to obtain and share due to license issues (Jimeno Yepes et al., 2013).

Recent work relied on a corpus of Cochrane Systematic Review Abstracts translated from English to French by professional translators at the French Cochrane Center. Preliminary work showed this dataset to be a good resource for domain-specific machine translation (Neveol et al., 2013). Follow-up work further developed the corpus using post-edited machine translation, which allowed the collection of a rich annotated parallel corpus(Ive et al., 2016)[3].

To encourage the community to take an interest in biomedical MT, recent challenges specifically provided targeted resources for system training and evaluation. The medical translation task at WMT 2014 (Bojar et al., 2014) included some parallel biomedical collections: MuchMore, various patents, Wikipedia titles, Khesmoi search queries, and vocabulary lists extracted from the UMLS. The biomedical track at WMT 2016 (Bojar et al., 2016) provided new resources for French, Portuguese and Spanish with the Scielo corpus (Neves et al., 2016). The biomedical track at WMT 2017 (Jimeno Yepes et al., 2017) continued to offer new resources for French (EDP corpus), Portuguese and Spanish (extension of the Scielo corpus). The task also relied on the UFAL corpus, which comprises some of the previously released sources (EMEA, patents, ...) as well as newly crawled online patient information covering a total of ten language pairs.

Previous work also reported comparable biomedical corpora collected from the Web for Spanish, Arabic and Japanese (Moreno-Sandoval and Campillos-Llanos, 2013). However, text seem to come from different sources and no study on the equivalence between the texts in the various languages seemed to have been carried out.

Table 1 presents a list of biomedical parallel corpora used in the literature. We provide information regarding the text genre and language pairs covered by each source as well as pointers to a description of the resources.

In the following sections, we describe more specifically the recent development of the EDP, ReBec and Scielo corpora.

[3] http://www.translatecochrane.fr/corpus/

3. Method for Corpus Development

For all corpora we produced, we carried out the following procedure: (a) document retrieval or download; (b) document parsing and processing; (c) document (sentence) alignment; and (d) quality checking.

Document retrieval and download. Document retrieval varies depending on the document collection, some are readily available for download while others need to be crawled from the corresponding Web site.

Document parsing and processing. After download, the documents need to be parsed in order to extract the relevant text, e.g., title and abstract in the case of scientific publications. Documents in the different languages are then paired based on identifiers in the source websites. This step depends on the format of the obtained documents, whether XML or HTML format. Finally, we split the document into sentences using standard NLP tools.

Document (sentence) alignment. The corpora described in this section do not result from organized professional translation. For this reason, the texts were not translated sentence by sentence as is often the case for professional translation of technical documents. Empirical inspection of the corpora suggests that while some of the documents reflect sentence by sentence translation, others were created more freely and the content in one language could be structured differently in the other language. We made the hypothesis that documents could nonetheless be aligned at the sentence level and we relied on automatic tools for performing the alignment. We identified alignment tools based on an evaluation of alignment for literary texts which is a genre that also features fuzzy alignment(Xu et al., 2015).

Quality checking. After automatically aligning the sentences of the documents, we manually checked a sample of our corpora. This was carried out using the Appraise (Federmann, 2010) tool, and we evaluated whether the aligned sentences where correct or whether more information was available in one language or the other. Native speakers of each foreign language were responsible for this task.

4. Application to Three Biomedical Corpora

Here we describe the three corpora that we developed and highlight the differences regarding the particular tools that we used for the various steps above.

EDP We identified five open access CC-BY journals, referenced EDP Sciences[4] as having content in French and in English: the articles were originally written in French but the journals also publish the titles and abstracts in English, using a translation provided by the authors. Three journals are listed by the publisher under *Health*: "Actualités Odonto-Stomatologiques" and "Médecine Buccale Chirurgie Buccale", which are journals addressing dentistry and "Les Cahiers de Myologie", a journal addressing muscle medicine. Two journals are listed under *Life & Environmental Sciences*: "Cahiers Agriculture" and "Oilseeds and fats, Crops and Lipids". A list of the journal URLs was ob-

[4] http://www.edpsciences.org

Corpus	genre	languages (other than English)	reference
Cochrane	Systematic Review (SR) abstracts	fr	(Ive et al., 2016)
COPPA, PaTr	Patents	de,fr	(Bojar et al., 2014)
EDP	Article titles and abstracts	fr	ibid.
EMEA	Medication description	cs,da,de,el,es,et,fi,fr,hu,it lt,lv,mt,nl,pl,pt,ro,sk,sl,sv	(Tiedemann, 2012)
Himl*	Patient information and SR abstracts	cs,de,fr	(Jimeno Yepes et al., 2017)
Khresmoi	Short medical search queries	cs,de,fr	(Pecina et al., 2013)
MEDLINE	Article titles	de,es,fr,hu,pl,tu	(Wu et al., 2011)
MuchMore Springer	Article titles and abstracts	de	(Widdows et al., 2002)
ReBEC	Clinical Trial summaries	pt	(Neves, 2017)
Santé Canada	Patient information	fr	(Deleger et al., 2009)
Scielo	Article titles and abstracts	es,fr,pt	(Neves et al., 2016)
UFAL*	Medical web crawl	cs,de,es,fr,hu,pl,ro,sv	(Jimeno Yepes et al., 2017)
UMLS	Metathesaurus	cs,da,de,el,es,eu, et,fi,fr,he,hu,hr it,ja,ko,lt,lv,nl,no,pl,pt,sv,tr,zh	(Lindberg et al., 1993)

Table 1: Overview of biomedical parallel corpus. We use ISO 639-1 two-letter language codes. A star indicates resources that include previously developed corpora as well as new data

Corpus	Tokens	Count method
EDP		
EN	56,684	`wc -w` on txt files
FR	62,333	`wc -w` on txt files
ReBEC		
EN	625,881	reported by (Neves, 2017)
PO	665,325	reported by (Neves, 2017)
Scielo		
EN	20,337,385	script `BioC2txtWithCounts.py`
ES	21,651,629	available on GitHub
EN	525,866	reported by (Neves et al., 2016)
FR	735,486	reported by (Neves et al., 2016)
EN	18,769,613	script `BioC2txtWithCounts.py`
PT	18,573,561	available on GitHub

Table 2: Content of open biomedical parallel corpus.

tained [5] and crawled [6] on March 15, 2017. The html pages were parsed to extract the titles and abstracts in French and English as well as the author names. Any articles lacking some of this information were discarded.

The dataset was pre-processed for sentence segmentation using the Stanford CoreNLP toolkit[7] for use in the WMT17 biomedical task. A manual reference for sentence segmentation was then created independently by revising baseline segmentation after the punctuation marks: full stop, interrogation point, exclamation point and colon.

Based on the manually validated sentence segmentation, the dataset was aligned automatically at the sentence level using YASA(Lamraoui and Langlais, 2013). Manual evaluation conducted on a sample set suggests that 94% of the sentences are correctly aligned, with about 20% of the sentence pairs exhibiting additional content in one of the languages.

MEDLINE vernacular titles MEDLINE indexes journals in languages other than English that publish a title and abstract in English. In this case, MEDLINE citations include the title of the article in the vernacular language in addition to English. This has been used to develop parallel corpora to train machine translation methods (Wu et al., 2011; Jimeno Yepes et al., 2013).

We have retrieved the MEDLINE citations for articles in French, Spanish and Portugese available before the first WMT biomedical task [8]. We collected titles in English and vernacular (Spanish, French and Portuguese). Titles are already aligned since they typically can be considered as one sentence. It can be noted that while our work was limited to the languages of interest in the WMT biomedical track, parallel titles and/or abstracts may also be retrieved for other languages. For instance, the query `chinese [la]` returns 286,151 results on September 29, 2017, and parsing the MEDLINE xml result file could yield several thousand aligned titles and abstract sentences for the relevant citations.

ReBEC As already described in (Neves, 2017), the construction of the ReBEC corpus followed the workflow described in the previous section. The Website site of the Brazilian Clinical Trials Registry [9] provides ways to easily download the trials in XML format, which was further

[5] Using `http://www.xsitemap.com/`

[6] Using the perl utility wget

[7] `https://stanfordnlp.github.io/CoreNLP/`

[8] using the pubmed queries `french [la]`, `spanish[la]`, `portugese [la]`.

[9] http://www.ensaiosclinicos.gov.br/

parsed. However, given the various elements (sub-sections) in a trial, e.g., inclusion criteria, exclusion criteria, and given that some of these appear multiple times in the document, the automatic alignment of parallel documents is not straightforward.

Scielo Scielo (Scientific Electronic Library Online) [10] is a database of open access scientific publications with a focus on developing and emerging countries, and especially on Latin America. All publications in Scielo are available under either the Creative Commons Attribution-Noncommercial 3.0 Unported (cc-by-nc) or Attribution 3.0 Unported (cc-by) licenses, which makes all documents suitable for redistribution and research purpose.

We developed a corpus based on Scielo (Scielo corpus (Neves et al., 2016)) using the following procedure. We crawled the Scielo site and retrieved articles periodically from Scielo. Our crawling has its starting point in the pages that list all journals from the "Biological Sciences" and "Health Sciences" subjects. These categories are used to compose the two datasets, with the corresponding names, of our corpus. Despite being distinct categories in Scielo, these are overlapping categories, as there are many journals that belong to both of them. From the list of journals, it is possible to retrieve a list of all issues of a particular journal, which is available in the regional web sites of Scielo in distinct countries, such as Brazil, Chile or Colombia. The HTML page of the journal's list of issues was further parsed to retrieve the page containing the list of articles of a given issue.

Finally, we downloaded the page of a particular article and parsed the HTML code in order to extract the title and the abstract of each publication. Titles and abstracts were subsequently stored and indexed in the SAP HANA database. All translations of the abstracts in Scielo are the original texts provided by the authors of the publications, who are presumably not professional translators, and who may not have native proficiency in both languages. After the initial version of the corpus produced in 2016, we are using the same procedure to update the corpus on a yearly basis for the ENES and ENPT language pairs. For ENFR, there were no new documents added in 2017.

5. Results

5.1. Datasets Descriptive Statistics

Table 2 presents detailed statistics of the contents of the biomedical parallel corpora that we developped. Table 3 presents an overview of the corpora with the training and test set splits that were offered thoroughout the WMT campaigns.

5.2. Quality Assessment

We also provide a summary of the correct alignment rate for the various corpora, as shown in Table 4. The alignment was automatically carried out using the respective tools as previously described and a sample was manually checked using the Appraise tools for ReBEC and Scielo, and manual inspection of text files for EDP.

[10] http://www.scielo.org

For EDP the manual reference for sentence segmentation provides an evaluation of Stanford sentence segmentation, which comes to 0.77 F-measure on the French portion and 0.81 F-measure on the English portion. Overall, error analysis reveals that the segmentation errors produced by the Stanford tool mainly result from segmentation of the section titles in structured abstracts (Introduction, Material and Method, Results...) which were considered as separate segments by the manual reference but not by the tools. Other errors occur due to organism names (e.g. `E. coli`, which may cause a sentence boundary to be set after "E.").

5.3. Data format

All corpora presented in the previous section are available from GitHub in the BioC format (Comeau et al., 2013), a standard XML format in the BioNLP community.

6. Discussion

In this section we present a short discussion on some interesting topics that raised during both the corpus construction and its use in our shared tasks.

6.1. Lessons learned during corpus construction.

The challenges of building parallel corpora for the biomedical include the identification of high quality relevant data that can be shared with the community. Technical issues then lie with the identification of adequate tools for sentence segmentation and alignment.

Sentence segmentation: we relied on tools which are non-specific for the biomedical domains, such as Stanford CoreNLP, OpenNLP and SAP HANA. However, we did observe issues. A specific discussion of sentence segmentation errors is reported in (Neves, 2017) for ReBEC. For the EDP corpus, we used intially used Stanford CoreNLP for sentence segmentation (in the version of the corpus distributed at WMT17). Then, we manually validated sentence segmentation in both languages in order to create a reference corpus that may be used to train and evaluate sentence segmentation tools. Therefore, Updated versions of the corpus reflect the manual sentence segmentation.

Sentence Alignment: GMA was used for Scielo and ReBEC. Due to difficulties to install GMA, Yasa(Lamraoui and Langlais, 2013) was used for EDP; however, Yasa may be limited to the language pair en/fr. We can refer readers to (Xu et al., 2015) for a discussion and evaluation of alignment tools for a specialized domain (literary texts). Nevertheless, both tools provided good automatic alignments (?). Additionally, GMA was used for two languages (es and pt) and two document types (scientific publications and clinical trials).

6.2. Differences across the corpora.

Despite the corpora presented in this work have been converted into a similar layout, we observed some differences across the results we obtained. These differences are mostly related to particularities of the corpora, such as its format. One such example is the lower rate of correct alignments for the ReBEC corpus that was due to problems when parsing the document format rather that the alignment tool itself, though some few errors could have come from GMA.

Training sets	fr/en	pt/en	es/en
MEDLINE	612,797/idem	74,286/idem	285,408/idem
Scielo	1,135/∼9,500	83,839/∼650,000	93,528/∼750,000
ReBEC		1188/∼23,000	

Test sets	fr/en	pt/en	es/en	en/fr	en/pt	en/es
Scielo	500/∼5,000	1000/∼8,000 189/1,897	1000/∼9,000 158/1,180	500/∼5,000	1,000/∼8,000 188/1,806	1000/∼9,000 158/1,082
EDP	85/699			84/750		

Table 3: Overview of the training and test sets. We present the number of documents and sentences in each corpus. Statistics for the MEDLINE dataset corresponds to both documents and sentences, given that it consists only of titles. For the Scielo test set collection, we present details of the WMT'16 (first row) and WMT'17 (second row) test sets.

Corpus	Rate of correct alignment
EDP	94%
ReBEC	67%
Scielo	79%-85%

Table 4: Summary of the rate of correct alignment for the corpora. No alignment was necessary for the MEDLINE titles.

6.3. Gaps to be addressed.

We can see from table 1 that there is no clinical corpus or datasets from the social media. Also, some languages benefit from better coverage than others depending on the corpora: DE, ES, FR, PT. Finally, we do not yet cover Asian languages, which we plan to address in the future through collaboration with members of the BioNLP community. Typically, we intend to create an abstract collection from MEDLINE as indicated in section 4.

7. Conclusion

We presented a scoping review of the various parallel corpora that are available for the biomedical domain. To our knowledge, this is the first survey of biomedical parallel corpora. In addition we detailed the development of corpora that we recently provided for training and evaluating biomedical machine translation systems. The collections cover a total of four languages (including English) and various types of documents, such as scientific publications and clinical trials, from various sources and databases. Further, these corpora have been evaluated on the scope of two shared tasks and are freely available for the scientific community either for MT or other NLP tasks. Finally, future work will contribute towards the inclusion of additional languages, e.g., German, as well as other documents types, e.g., health-related news and clinical reports.

8. Acknowledgements

This work was supported in part by the Agence Nationale pour la Recherche (French National Research Agency) under grant number ANR-13-JCJC-SIMI2-CABeRneT. The authors would like to thank Arthur Boyer for his contribution to the EDP corpus.

9. Bibliographical References

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation (WMT16) at the Conference of the Association of Computational Linguistics*, pages 131–198. 0.

Bueno, F. T. C. (2017). Vigilância e resposta em saúde no plano regional: um estudo preliminar do caso da febre do Zika vírus. *Ciêcia & Saúde Coletiva*, 22:2305 – 2314, 07.

Chiao, Y. and Zweigenbaum, P. (2004). Aligning words in french-english non-parallel medical texts: effect of term frequency distributions. In *Stud Health Technol Inform*, volume 107, pages 23–7.

Comeau, D. C., Islamaj Doğan, R., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T. C., Wu, C. H., and Wilbur, W. J. (2013). Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013:bat064.

Deleger, L., Merkel, M., and Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *J Biomed Inform*, 42(4):692–701, Aug.

Federmann, C. (2010). Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *In LREC*.

Hellrich, J., Clematide, S., Hahn, U., and Rebholz-Schuhmann, D. (2014). Collaboratively annotating multilingual parallel corpora in the biomedical domain - some mantras. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ive, J., Max, A., Yvon, F., and Ravaud, P. (2016). Diagnosing high-quality statistical machine translation using traces of post-edition operations. In *Proceedings of the LREC 2016 Workshop: Translation evaluation - From*

fragmented tools and data sets to an integrated ecosystem, pages 55–62, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Jimeno Yepes, A. and Névéol, A. (2013). Effect of additional in-domain parallel corpora in biomedical statistical machine translation. In *Proceedings of the 4th International Workshop on Health Document Text Mining and Information Analysis with the Focus of Cross-Language Evaluation (Louhi 2013)*.

Jimeno Yepes, A., Prieur-Gaston, E., and Neveol, A. (2013). Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14(1):146.

Jimeno Yepes, A., Névéol, A., Neves, M., Verspoor, K., Bojar, O., Boyer, A., Grozea, C., Haddow, B., Kittner, M., Lichtblau, Y., Pecina, P., Roller, R., Rosa, R., Siu, A., Thomas, P., and Trescher, S. (2017). Findings of the WMT 2017 Biomedical Translation Shared Task. In *Proceedings of the Second Conference on Machine Translation (WMT17) at the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Copenhagen, Denmark, September.

Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M., and Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Lamraoui, F. and Langlais, P. (2013). Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment? In *XIV Machine Translation Summit*, Nice, France, Sept.

Lindberg, D., Humphreys, B., and McCray, A. (1993). The unified medical language system. *Yearb Med Inform*, 1:41–51.

Liu, H., Christiansen, T., Baumgartner, W. A., and Verspoor, K. (2012). Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3(1):3, Apr.

Moreno-Sandoval, A. and Campillos-Llanos, L. (2013). Design and annotation of multimedica - a multilingual text corpus of the biomedical domain. *Procedia - Social and Behavioral Sciences*, 95:33 – 39. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Neveol, A., Max, A., Ivanishcheva, Y., Ravaud, P., Zweigenbaum, P., and Yvon, F. (2013). Statistical machine translation of systematic reviews into French. In *Proc Workshop on optimizing understanding in multilingual hospital encounters – TIA 2013*, pages 10–13.

Neves, M., Jimeno Yepes, A., and Névéol, A. (2016). The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. 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).

Neves, M. (2017). A parallel collection of clinical trials in portuguese and english. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 36–40, Vancouver, Canada, August. Association for Computational Linguistics.

Ozdowska, S., Neveol, A., and Thirion, B. (2005). Traduction compositionnelle automatique de bitermes dans des corpus anglais/franÃ§ais alignÃ©s. In *Proceedings of the 6th International Conference on Terminology and Artificial Intelligence*, pages 83–94.

Pecina, P., Dušek, O., Hajič, J., and Urešová, Z. (2013). Khresmoi query translation test data 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

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'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Walker, C. (2016). Learn the local lingo to get ahead. *Nature*, 7607(534):425–7, Jun 16.

Widdows, D., Dorow, B., and Chan, C.-K. (2002). Using parallel corpora to enrich multilingual lexical resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Wu, C., Xia, F., Deleger, L., and Solti, I. (2011). Statistical machine translation for biomedical text: are we there yet? In *Proc AMIA Annu Symp*, pages 1290–9, Nov.

Xu, Y., Max, A., and Yvon, F. (2015). Sentence alignment for literary texts. *Linguistic Issues in Language Technology*, 12(6), Oct.

Zeng-Treitler, Q., Kim, H., Rosemblat, G., and Keselman, A. (2010). Can multilingual machine translation help make medical record content more comprehensible to patients? In *Stud Health Technol Inform*, volume 160, pages 73–7.