Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1068–1072
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Constructing Parallel Corpora from COVID-19 News using MediSys Metadata

Dimitris Roussis, Vassilis Papavassiliou, Sokratis Sofianopoulos, Prokopis Prokopidis, Stelios Piperidis
Institute for Language and Speech Processing (ILSP)
Athena Research and Innovation Center, Athens, Greece
{dimitris.roussis, vpapa, s_sofian, prokopis, spip}@athenarc.gr

Abstract

This paper presents a collection of parallel corpora generated by exploiting the COVID-19 related dataset of metadata created with the Europe Media Monitor (EMM) / Medical Information System (MediSys) processing chain of news articles. We describe how we constructed comparable monolingual corpora of news articles related to the current pandemic and used them to mine about 11.2 million segment alignments in 26 EN-X language pairs, covering most official EU languages plus Albanian, Arabic, Icelandic, Macedonian, and Norwegian. Subsets of this collection have been used in shared tasks (e.g. Multilingual Semantic Search, Machine Translation) aimed at accelerating the creation of resources and tools needed to facilitate access to information in the COVID-19 emergency situation.

Keywords: COVID-19, Parallel Corpus, Machine Translation, Emergency Situations

1. Introduction

In December 2019, the first cases of the COVID-19 disease caused by the SARS-CoV-2 virus were identified in Wuhan, Hubei, China. On 11 March 2020, the WHO (World Health Organization) declared the novel coronavirus outbreak a pandemic and it was only a matter of time before every country in the world would be affected by it to a smaller or larger extent. The COVID-19 pandemic, however, unfolded as something more than a disease; it became the catalyst of -quite simultaneous- global changes in everyday life and public discourse with a tremendous impact on financial stability, healthcare systems, national policies, working conditions, individual freedoms, travel and movement.

Due to the pandemic, we became increasingly dependent on mass media, social media, news websites, online communication services and in general, anything that involved internet and the "digital". Thus, an abundance of COVID-19 related textual information was generated in many languages and platforms, and motivated researchers to gather and structure data in order to facilitate the use of statistical methods and Artificial Intelligence (AI) for a plethora of applications (Shuja et al., 2020). The emergence of multidisciplinary initiatives, such as the creation of the CORD-19 (Covid-19 Open Research Dataset) dataset[1] and the shared tasks built around it, illustrated how data scientists, biomedical experts and policy makers can connect and cooperate in the battle against COVID-19 (Wang et al., 2020).

In emergency situations, Machine Translation (MT) can be embedded in a rapid-response infrastructure and accelerate the speed in which relief can be provided (Lewis et al., 2011). The COVID-19 pandemic is such an emergency situation, albeit with different characteristics compared to other crises due to its global character and extended duration. Access to guidelines, news, announcements, high-quality articles etc. is vital to a wide range of stakeholders, from healthcare professionals, decision-makers and researchers to vulnerable groups, such as immigrants, marginalized communities and the oldest parts of the population. Therefore, MT plays a crucial role in information dissemination and triage, as well as fact-checking, especially if we take into account the existence of the "infodemic"[2], i.e. the unprecedented scale of misinformation campaigns, "fake news" and conspiracy theories (propagated even by public figures) during the pandemic.

This paper discusses a use case of exploiting the outcomes of an available infrastructure, namely MediSys (Medical Information System), with the purpose of constructing parallel corpora beneficial for MT. It could be seen as an application of simulating a quick response of the MT community to the pandemic crisis. MediSys is one of the publicly accessible systems of the Europe Media Monitor (EMM) which processes media to identify potential public health threats in a fully automated fashion (Linge et al., 2010). A dataset of metadata which concerns COVID-19 related news was made publicly available and contains millions of news articles stored in RSS/XML format (Jacquet et al., 2020). We parsed these metadata to extract URLs and created monolingual news corpora by harvesting them. Considering these corpora comparable, since they are focused on COVID-19 and were based on news articles which were published during the same period, we processed them with the purpose of identifying parallel segments and building corpora suitable for training and adapting MT systems.

It is worth mentioning that some of the constructed corpora have already been used (alongside other sources) at the COVID-19 MLIA-Eval[3], an initiative focused on COVID-19 related multilingual information access with specific evaluation tasks open to participants. In particular, the monolingual corpora were used for the Multilingual Semantic Search task while the parallel corpora became available for the MT task. More information on this task can be found in Casacuberta et al. (2021).

[1] https://www.semanticscholar.org/cord19
[2] https://www.who.int/health-topics/infodemic
[3] http://eval.covid19-mlia.eu/

2. Related Work

Although there is a plethora of grand-scale multilingual corpora that can be utilized in Machine Translation systems, the pandemic ignited interest in the rapid translation of COVID-19 related information and, subsequently, various corpora, benchmarks, initiatives and expert translation systems emerged.

TICO-19 (Translation Initiative for COvid-19) is a large collaborative project which has published validation and test sets and resources related to COVID-19 in several languages that act as a benchmark for Machine Translation systems (Anastasopoulos, 2020). For instance, the ParaCrawl project released a synthesized data corpus[4] focused on the COVID-19 domain, which utilizes parallel sentences from available parallel corpora in combination with a COVID-19 glossary sourced from TICO-19. TAUS launched the Corona Crisis Corpus project as a collective industry charity effort, which made available parallel data[5] related to virology, epidemic, medicine, and healthcare. SYSTRAN contributed to this initiative by producing Corona Crisis Translation Models[6].

The 4th LoResMT (technologies for machine translation of low resource languages) workshop introduced a shared task focused on texts and sign language specifically related to COVID-19 (Ojha et al., 2021). The datasets provided by the organizers have been mainly gathered from governmental sources and news sites, as well as press conferences for the COVID-19 response. Similarly, the 1st round of the COVID-19 MLIA-Eval initiative provided an updated version of EMEA corpus[7] and web-acquired data from websites of national authorities and public health agencies[8], EU agencies and specific broadcast portals.

There are also several applications that aim at constructing or adapting Machine Translation systems for the accurate translation of COVID-19 related texts. For example, in Way et al. (2020), eight Neural Machine Translation (NMT) systems (for high-resource language pairs) were developed with the use of relevant datasets, such as the TAUS Corona Crisis corpus and the OPUS EMEA[9] corpus. Other approaches include Mahdieh et al. (2020), where the authors report significant quality improvements by adapting an existing NMT system using monolingual in-domain data (e.g. English data related to COVID-19) as well as the one in Bandyopadhyay (2020), where the focus lies on translating five low resource language pairs using factored NMT (i.e. NMT with additional linguistic features in the English side of the sentence pairs).

3. Methodology

In this section we describe the method that we used, which includes fetching news articles, merging them into monolingual corpora, mining parallel sentences and applying filtering techniques to construct parallel corpora in 26 language pairs.

The first step of our pipeline was to parse the MediSys metadata[10]. Originally, the MediSys dataset comprised of automatically extracted metadata (e.g. URL, title, language, identified named entities, etc.) from news articles related to COVID-19. However, we only utilized the language and URL of each entry; in particular, we selected datasets spanning across 10 months (December 2019 to September 2020) and located the articles of several languages[11] summing up to about 57 million URLs. Their distribution according to the targeted languages is illustrated in Figure 1. Most of the articles were in English, Spanish, Italian and Arabic (in descending order).

The web pages were fetched as HTMLs using the GNU Wget package[12], and the boilerplate (e.g., advertisements, disclaimers, etc.) was removed to keep the main content of each webpage in plain text format by using the Cleaner module of the ILSP-FC toolkit[13]. During data acquisition, we noticed that about 12% of the requested web pages were not fetched as the links were already deleted or moved.

MediSys categorizes the metadata into groups according to the publication date of the news articles i.e., in a month or two-month period. We adopted the same approach and grouped the resulting text files accordingly, e.g., batch of files for March or for April to May, etc. We merged the text files to create a single document for each language and each period, e.g. English news of March, Greek news of March, …, English news of April and May, Greek news of April and May, etc. These documents are the initial versions of the COVID-19 related monolingual corpora and could be considered comparable (in pairs), due to their narrow topic and the fact that they were published in the same time period. Especially during this universal emergency, many news articles which were published on popular broadcast portals, were also promptly republished on several portals in their native/local language. This motivated us to mine these corpora (i.e. pairs of comparable documents) for parallel segments.

To this end, each language-period document was split into sentences by using in-house tools based on the NLTK library[14] (Bird et al., 2009), duplicates were detected and removed, and language identification[15] was applied. In total, approximately 150 million sentences remained and their distribution across the 29 used languages is presented in Figure 2. Unfortunately, very few sentences remained

[4] https://paracrawl.eu/manufactured-data

[5] https://md.taus.net/corona

[6] https://www.systransoft.com/systran/news-and-events/specialized-corona-crisis-corpus-models/

[7] Updated version of the EMEA corpus on ELRC-SHARE repository.

[8] https://www.ecdc.europa.eu/en/covid-19/external-resources

[9] https://opus.nlpl.eu/EMEA.php

[10] https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/LANGUAGE-TECHNOLOGY/EMM_collection/2020_MediSys_Covid19_dataset/

[11] EN (English), BG (Bulgarian), CS (Czech), DA (Danish), DE (German), EL (Greek), ES (Spanish), ET (Estonian), FI (Finnish), FR (French) , GA (Irish), HR (Croatian), HU (Hungarian), IS (Icelandic), IT (Italian), LT (Lithuanian), LV (Latvian), MT (Maltese), NL (Dutch), NB (Norwegian Bokmål), PL (Polish), PT (Portuguese), RO (Romanian), SK (Slovak), SL (Slovenian), SV (Swedish), MK (Macedonian), SQ (Albanian) & AR (Arabic)

[12] https://www.gnu.org/software/wget/

[13] http://nlp.ilsp.gr/redmine/projects/ilsp-fc

[14] https://www.nltk.org/

[15] https://github.com/shuyo/language-detection

for Maltese and Gaelic and consequently, these languages were omitted in the next steps.

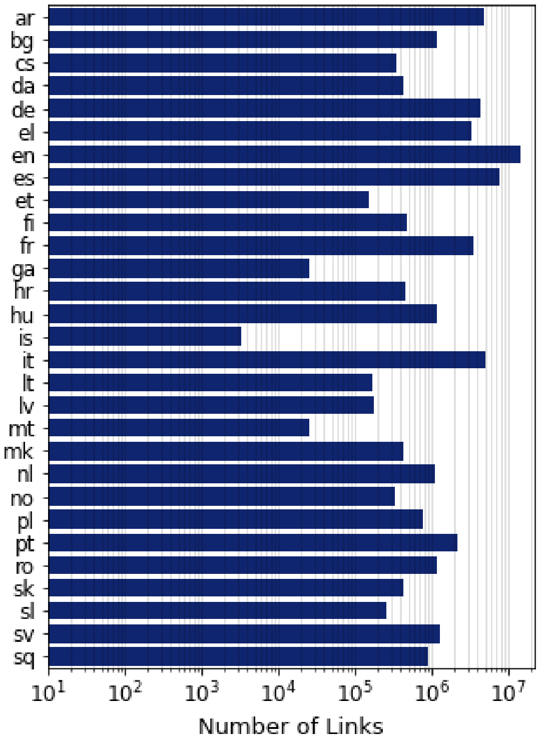

Figure 1: Total number of links per language in MediSys.

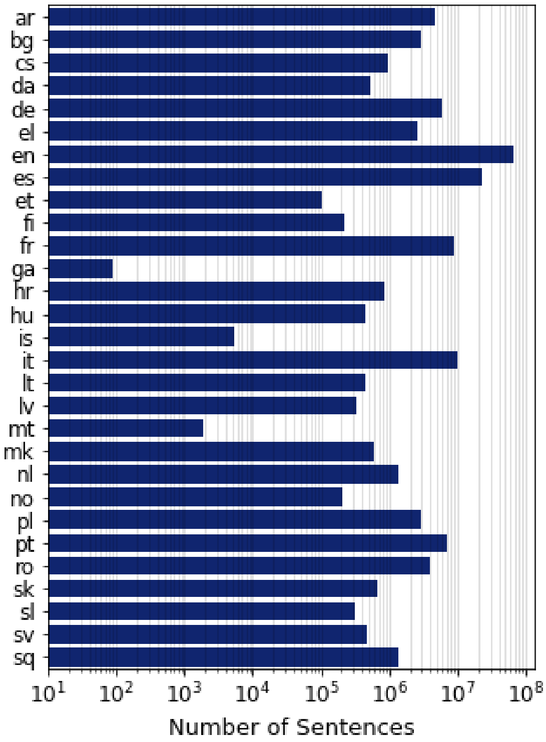

Figure 2: Total number of sentences per language.

Afterwards, the LASER toolkit[16] was applied on each document pair to mine sentence alignments for each EN-X language pair. This toolkit uses a pre-trained multilingual sentence encoder which maps sentences to fixed-length vectors in a common space and, for each source sentence, the k nearest neighbor (we use the default settings: k=4) target sentences are determined in terms of the cosine similarity between the source and the target vectors; and vice versa for each target sentence. Then, it uses a margin-based technique (Artetxe and Schwenk, 2019) for selecting the best sentence pair and penalizes sentences with many paraphrases, (i.e. cases in which a source sentence was paired with multiple target sentences as candidate translations and vice versa). Due to hardware constraints, we decided to divide English corpora (the largest ones) into smaller, more manageable chunks. On one hand, this approach allows us to overcome computational and memory-related issues. On the other hand, dividing the documents into chunks before utilizing LASER, may introduce incorrect sentence pairs as a given sentence in language L1 is used as an alignment candidate in each chunk of L2.

Taking this into account and having in mind the nature of our data i.e., news articles that might have been (partially) translated many times and republished on several websites, we examined the existence of paraphrases in the resulting collections of sentence pairs. We investigated cases in the EN-EL (English-Greek) corpus, in which the English part was the same whereas there were various paired EL segments. Experimentally, we concluded that long sentence pairs with a higher score tend to be valid alternative translations. Two indicative examples for EN-NL (English-Dutch and EN-IT (English-Italian) are presented in Table 1. We aim to preserve such cases while discarding the non-valid alternative translations, as they could prove useful for MT systems or in the construction of paraphrase datasets as in Hu et al. (2019).

Source Sentence	**Target Sentence**
T cells might be able to recognize virally infected cells and destroy them, limiting the virus's spread in the body.	T-cellen herkennen met virussen geïnfecteerde cellen en vertellen deze cellen dat ze zichzelf moeten vernietigen, zodat het virus zich niet door het hele lichaam kan verspreiden.
T-cells recognize virus-infected cells and tell those cells to self-destruct, preventing the virus from spreading throughout the body.	
Study co-author Matthew Gidden from Climate Analytics-Berlin said, "The lasting effect of Covid-19 on climate will not depend on what happens during the crisis, but what comes after.	Il coautore dello studio Matthew Gidden di Climate Analytics, è d'accordo: «L'effetto duraturo del Covid-19 sul clima non dipenderà da ciò che è accaduto durante la crisi, ma da ciò che verrà dopo.
"The lasting effect of COVID-19 on climate will not depend on what happens during the crisis, but what comes after," said study co-author Matthew Gidden.	

Table 1: Examples of alternative translations EN-NL (Upper 2 rows) EN-IT (Lower 2 rows)

[16] https://github.com/facebookresearch/LASER

The parallel data which resulted for each period were concatenated to form a single bilingual corpus per language pair. We applied the following filters in order to improve the quality and MT-readiness of the dataset:

- **Global Threshold:** Sentence pairs with a score lower than 1.04 were removed, as proposed in WikiMatrix (Schwenk et al. 2021).
- **Near Deduplication Filter:** Removes near duplicate sentence pairs which emerged because parallel data from different periods were merged into the same file for each language pair.
- **Maximum Sentence Length Filter:** Sentence pairs in which either the source or target sentence was longer than 80 tokens were removed.
- **Digit Mismatch Filter:** Removes sentence pairs in which the source and the target sentence have varying sequences of digits. Since this is a strict filter, we decided to apply it only to sentence pairs with a score below the threshold of 1.12.
- **Alternative Translations Filter:** As explained above, the bilingual corpora contain candidate alternative translations. To exclude the non-valid cases, we kept the sentence pairs with more than 10 tokens in both sentences and with an alignment score higher than 1.06.

4. Dataset Information

Overall, the final dataset comprises 11,248,573 sentence pairs in 26 EN-X language pairs, as presented in Table 2, and has been created in TMX (Translation Memory Exchange) and TSV (Tab-Separated Value) file formats. This COVID-19 related parallel resource covers 22 out of 24 official EU languages, as well as Albanian, Arabic, Icelandic, Macedonian and Norwegian.

There are important variations in the proportion of the parallel sentences among language pairs. For example, the EN-IS (English-Icelandic) corpus contains only a few thousand alignments, while the EN-ES (English-Spanish) corpus approximately 1.5 million.

The contents of the resource are related to COVID-19 and originated mainly from news articles collected from the MediSys dataset. Nevertheless, to further demonstrate the relatedness to COVID-19 of the final corpus, a word cloud of the approximately 60 million English sentences is presented in Figure 3. The word cloud was created using the Python Wordcloud[17] package using the sentence-splitted and lowercased English monolingual corpus after deduplication and removal of stop words (as well as the removal of the words "said" and "will"). As expected, COVID-19 related terms such as "coronavirus", "health", "covid", "pandemic", "medical", "virus", "outbreak", "lockdown" etc. occur quite frequently in the corpus. Furthermore, the frequent appearance of terms such as "global", "public", "state", "business", "national", "president", "market", "government", etc. is quite indicative of the effect of the pandemic on state governance and the global economy.

The news articles originated from several copyrighted websites. We are working towards resolving the IPR (Intellectual Property Rights) issues with the aim of making the resource publicly available.

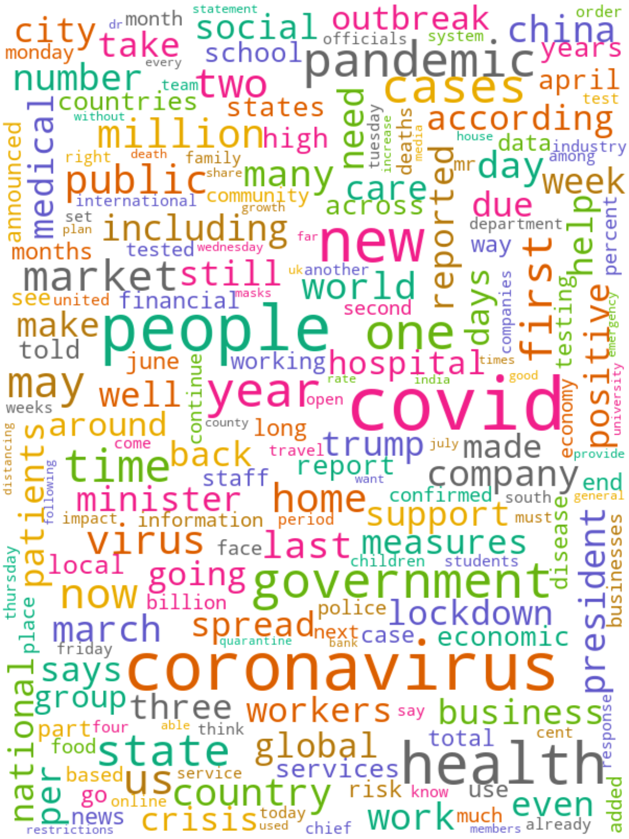

Figure 3: Word cloud of English monolingual corpus.

5. Conclusion and Future Research

In this paper, we described a collection of bilingual corpora which are related to COVID-19 news. Comparable monolingual corpora were constructed based on the metadata provided by the MediSys dataset and parallel sentences were extracted. Initially motivated by the fact that part of the dataset was used in the Machine Translation task of the COVID-19 MLIA @ Eval initiative, filtering criteria were applied so as to ensure that the bilingual corpora are ready for use in MT systems.

The dataset contains more than 11 million COVID-19 related sentence pairs in 26 EN-X language pairs and may prove a timely and useful resource in adapting MT systems which can be part of a response to the COVID-19 emergency or future similar crises.

In future work, we aim to further augment the resource with newly published content, given that batches of metadata for November and December 2021 are already available and motivated by the fact that the need to respond to the pandemic has not diminished.

Furthermore, we plan to mine segment alignments across more language pairs (which do not include English) and construct multilingual parallel corpora, as well as to assess the potential of the resource for adapting NMT systems to the domain of COVID-19.

[17] https://pypi.org/project/wordcloud/

Language Pair	Number of Sentence Pairs
en-ar	355,536
en-bg	662,595
en-cs	244,503
en-da	171,727
en-de	1,076,666
en-el	529,518
en-es	1,488,765
en-et	70,803
en-fi	111,589
en-fr	1,134,809
en-hr	230,857
en-hu	111,992
en-is	4,265
en-it	738,917
en-lt	149,897
en-lv	123,893
en-mk	182,438
en-nl	462,361
en-no	111,642
en-pl	565,915
en-pt	1,062,473
en-ro	693,614
en-sk	207,286
en-sl	122,186
en-sq	344,204
en-sv	283,122
Total	**11,248,573**

Table 2: Number of sentence pairs per language pair

6. Acknowledgements

This work has been supported by the European Language Resource Coordination (ELRC), a service contract under the EC's Connecting Europe Facility SMART 2019/1083 programme. We thank Guillaume Jacquet for his support in handling MediSys dataset of metadata. We would like to thank COVID-19 MLIA members who helped in data acquisition. Access to the Okeanos cloud service[18] provided by the Greek Research & Technology Network, is greatly appreciated.

7. Bibliographical References

Anastasopoulos, A., Cattelan, A., Dou, Z.Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P. and Lazar, R. (2020). TICO-19: the Translation Initiative for Covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Virtual Conference, November. Association for Computational Linguistics (ACL).

Artetxe, M. and Schwenk, H. (2019). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197-3203, Florence, Italy, July. Association for Computational Linguistics (ACL).

Bandyopadhyay, S. (2020). Factored Neural Machine Translation on Low Resource Languages in the COVID-19 crisis.

Bird, S., Klein, E. and Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Casacuberta, F., Ceausu, A., Choukri, K., Deligiannis, M., Domingo, M., Garcıa-Martınez, M., Herranz, M., Papavassiliou, V., Piperidis, S., Prokopidis, P. and Roussis, D. (2021). The Covid-19 MLIA@ Eval Iniative: Overview of the machine translation task.

Hu, J. E., Rudinger, R., Post, M. and Van Durme, B. (2019). ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pages 6521-6528.

Jacquet, G. and Verile, M. (2020). COVID-19 news monitoring with Medical Information System (Medisys). *European Commission, Joint Research Centre (JRC)*.

Lewis, W. D., Munro, R. and Vogel, S. (2011). Crisis MT: Developing A Cookbook for MT in Crisis Situations. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 501–511.

Linge, J.P., Steinberger, R., Fuart, F., Bucci, S., Belyaeva, J., Gemo, M., Al-Khudhairy, D., Yangarber, R. and van der Goot, E. (2010). MediSys: medical information system. In *Advanced ICTs for disaster management and threat detection: Collaborative and distributed frameworks*, pages 131-142, IGI Global.

Mahdieh, M., Chen, M. X., Cao, Y. and Firat, O. (2020) Rapid Domain Adaptation for Machine Translation with Monolingual Data. *arXiv preprint arXiv:2010.12652*.

Ojha, A.K., Liu, C.H., Kann, K., Ortega, J., Shatam, S. and Fransen, T. (2021). Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-resource Languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 114-123, August.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H. and Guzmán, F. (2021). WikiMatrix: Mining 135m Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Main Volume, pages 1351-1361.

Shuja, J., Alanazi, E., Alasmary, W., and Alashaikh, A. (2020). COVID-19 open source data sets: a comprehensive survey. *Applied Intelligence*, 1-30.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R. M. and Li, Y. (2020). CORD-19: The Covid-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Way, A., Haque, R., Xie, G., Gaspari, F., Popović, M. and Poncelas, A. (2020). Rapid Development of Competitive Translation Engines for Access to Multilingual COVID-19 Information. *Informatics*, 7(2):19.

[18] https://okeanos.grnet.gr/