COVID-19 Mythbusters in World Languages

Mana Ashida1, Jin-Dong Kim2, Seunghun J. Lee3,4
1 Yahoo Japan Corporation, 2 Database Center for Life Science, ROIS, 3 International Christian University, 4 IITG
maashida@yahoo-corp.jp, jdkim@dbcls.rois.ac.jp, seunghun@icu.ac.jp

Abstract
This paper introduces a multi-lingual database containing translated texts of COVID-19 mythbusters. The database has translations into 115 languages as well as the original English texts, of which the original texts are published by World Health Organization (WHO). This paper then presents preliminary analyses on latin-alphabet-based texts to see the potential of the database as a resource for multilingual linguistic analyses. The analyses on latin-alphabet-based texts gave interesting insights into the resource. While the amount of translated texts in each language was small, character bi-grams with normalization (lowercasing and removal of diacritics) turned out to be an effective proxy for measuring the similarity of the languages, and the affinity ranking of language pairs could be obtained. Additionally, the hierarchical clustering analysis is performed using the character bigram overlap ratio of every possible pair of languages. The result shows the cluster of Germanic languages, Romance languages, and Southern Bantu languages. In sum, the multilingual database not only offers fixed set of materials in numerous languages, but also serves as a preliminary tool to identify the language family using text-based similarity measure of bigram overlap ratio.

Keywords: COVID-19 Mythbusters, cross-linguistic database, clustering analysis

1. Introduction
The rapid progress of natural language processing (NLP) applications are often limited to languages that already have multitude of resources such as English, French or Japanese. As such, a significant number of languages do not have any resources for NLP applications [Joshi et al., 2020]. Such a disparity between languages turned out to be problematic under the COVID-19 pandemic situation where information sharing became important all over the world. Additionally, preventing the spreading of misinformation became equally important. The lack of NLP applications that target languages with low resources prevented necessary health measures from circulating promptly to regions where these languages are spoken.

At the onset of the COVID-19 pandemic, the World Health Organization (WHO) recognized problems that arise due to misinformation that circulated via Social Networking Sites or Applications. Soon, a webpage that aims to bust the myths concerning COVID-19 was created to raise awareness of these false beliefs that were spreading online [Lee and Won, 2021]. This paper has two aims. First, it reports the creation of a freely downloadable web-based database in about 116 versions, which was created based on the initiative “COVID-19 Mythbusters in World Languages” created by the third author. The 115 languages include 40 languages that do not have a single page on wikipedia which is written in these languages. As of January 14, 2022, Wikipedia is available in 341 language. Second, results of preliminary analyses of the latin-based texts in the database is reported. We calculate bigram overlap ratio between two languages using several normalization strategies such as lowercasing and removing diacritics to examine which normalization yield the better identification of related language pairs. Our results show that removing diacritics and replacing capital letters to lower case letters lowered the number of unexpected language pairs, whereas removing space increased the possibility of odd language pairs, suggesting space between words functions as an important delimiter in written forms in all these languages. Moreover, we examine whether a clustering analysis based on the character bigram overlap ratio of every possible language pair can identify genetically related languages.

2. Background
2.1. COVID-19 Related Language Resource Creation
During the COVID-19 pandemic, multi-lingual projects emerged. A team of researchers at the social center of Oxford University hosted a project on parenting during pandemic based on information available on the WHO website. The parenting tips are now translated into over 100 languages [2]. Translations without borders launched a COVID-19 Community Translation Program to assist communities that need help with translating COVID-related information. The project currently has translators available in 106 language [3]. Endangered languages Fund has a resource website for languages that are indigenous, endangered or under-resourced [4].

https://www.covid19parenting.com/#/
https://translatorswithoutborders.org/translations-covid-19/
https://endangeredlanguagesproject.github.io/COVID-19/
2.2. Participatory Research on Low-Resource Languages

Participatory Research involves researchers and communities. It aims the transition from research to action through democratization of science, and values the benefit of communities (English et al., 2018). The applicability of participatory research has recently been tested on language resource creation process.

Nekoto et al. (2020) is a case study of participatory research on machine-translation in African Languages, led by a community “Masakhane”, which highlighted the importance of the resources for machine translation (MT) systems built and evaluated by the people who speak and use the target languages. Nekoto et al. (2020) also show that participatory research can benefit the low-resourced MT development.

Our project is in line with participatory research on low-resource languages in terms of valuing the benefit of the community. The translations have been produced by translating volunteers most of whom speak the languages as their first language. The background of the volunteers were diverse ranging from students and community linguists to university professors or professional translators. For some languages, resources in our project is the only digitized material that can be found on the Internet. For example, 40 out of 115 languages do not have a single wikipedia page. While our database is too small to create a full-fledged machine translation tool, we expect that the database can be used as a test set to evaluate existing machine translation algorithms, especially the ones used in the medical and health domains.

3. Project Overview

3.1. Mythbusters in World Languages

The main website offers freely available resources that are translations of COVID-19 mythbusters that are originally compiled by the World Health Organization (WHO). The original edition is “Coronavirus disease (COVID-19) advice for the public: Mythbusters. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO”. The first version of the texts used the 19 mythbusters that were available in early April, 2020. A later version used 25 mythbusters that was extended in summer 2020. The first release of the website was in mid April when translations of 9 languages (including English and Japanese) became available. Picture panels that are adjustable to smart phone screens were produced with translated texts and icons corresponding to the meaning of a myth buster. Thereafter, additional languages were added to the website, which now lists 116 versions including Japanese Sign Language.

Our project is in line with participatory research on low-resource languages in terms of valuing the benefit of the community. The translations have been produced by translating volunteers most of whom speak the languages as their first language. The background of the volunteers were diverse ranging from students and community linguists to university professors or professional translators. For some languages, resources in our project is the only digitized material that can be found on the Internet. For example, 40 out of 115 languages do not have a single wikipedia page. While our database is too small to create a full-fledged machine translation tool, we expect that the database can be used as a test set to evaluate existing machine translation algorithms, especially the ones used in the medical and health domains.

The database consists of three tables: (a) Translations, (b) Languages, and (c) Scripts. The Translations table contains all the translations of the COVID-19 mythbusters each of which is associated with a language.
and a script. The Languages table contains all the languages, and the Script table contains all the scripts.

Three languages (Mala Pulaya, Serbian, Yiddish) have two versions of the mythbusters because the language can be written in more than one script. Mala Pulaya is spoken in the state of Kerala, India, and it is written in Malayalam or Tamil. Mala pulaya speakers living in the Malayalam area read the Malayalam script, but Mala pulaya speakers in the Tamil area are literated in the Tamil script. Serbian is written in Latin alphabet or in Cyrillic alphabet. Yiddish is written in Hebrew or in Latin alphabet. For the reason, the three languages, Mala Pulaya, Serbian and Yiddish are associated with two different versions of translations for each entry.

Figure 1 shows a screenshot of the front page of the database homepage, which shows a language section area in the middle. A click of a language name in the language selection area leads to a page that shows the translations of the mythbusters as in Figure 2. In the original mythbusters published by WHO, each mythbuster passage has its own serial number, which is retained in all the translations of the passage. In other words, all the translations of a mythbuster passage are associated through the serial number.

Two types of csv files are available for downloading and for further analyses of the data. The metadata file has a list of language names, script names and the names of the contributors (i.e. translators). The “Download translations” button generate a csv file with all the translations of the mythbusters. All characters in the translations are encoded in UTF-8. The website also features a function that allows registered users to add translations of a new language. Prospective contributors may email the webmaster or directly fill Google forms using links available on the website.

4. Exploratory Data Analysis

As a multi-lingual parallel corpus, the dataset can be characterized as (1) small, (2) specific (focused), and (3) broad. Including only 18 to 24 sentences (or paragraphs) per language, it is certainly a small corpus, and the subject domain of the content is highly focused. However, the real value of the data set is in the diversity of the languages which include substantial number of truly endangered languages, apart from its pragmatic value that it helps the spread of critical information to the people of under-represented languages.

Releasing the dataset for the community of NLP, we hope the dataset to be useful for developing NLP tools or resources for particularly low-resource languages. Although the size of the texts per language is small, potential usage we look forward to seeing include discovery of evolutionary relationship of the languages, and furthermore application of recent technologies such as transfer learning or few-shot learning which leverage much richer resources in relevant other languages for development of practically useful NLP tools. Toward that direction, we report the results of our preliminary data analyses, which are designed to see whether the dataset contains sufficient amount of signal to measure similarity and differences of the languages.

4.1. Language Taxonomy Based on Character Overlap

Our preliminary analyses were performed with the surface form of the script of the languages. Our database has a mixture of scripts that include the latin alphabet, the greek alphabet, the cyrillic alphabet, Hangul, the Arabic script, various syllabaries (Japanese, Thai, Burmese, Tibetan, Nuosu Yi, Hindi, Tamil, Malayalam, Telugu etc.), and logographs such as Chinese characters. We focused on only the languages which use Latin alphabets, because they form the largest group in terms of the scripts, and because the other languages use scripts whose surface form is completely different from Latin alphabets. The Latin alphabet group includes 65 languages.

A mythbuster in the database is a sentence or a paragraph. The average number of words per mythbuster ranges from 8.83 (Xhosa) to 30.33 (Kiribati), and Appendix B shows the descriptive data of all the languages targeted in this paper. The five most frequent words in each language are also listed, but any word that has less than five characters is excluded to avoid ending up with a list of function words.

For representation of the texts for the analyses, two set-based text modeling methods, the bag-of-word-bigrams
model and the bag-of-character-bigrams model, and three text normalization strategies, lowercasing, removal of diacritics, and removal of inter-words spaces, were tested.

For the similarity measure between any pair of texts, the Jaccard similarity coefficient is adopted, which is calculated as Equation 1, where $T_1$ and $T_2$ are the pair of texts which are modeled as either bag-of-word-bigrams or bag-of-character-bigrams.

$$\text{Jaccard}(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \quad (1)$$

Hereafter, we call the Jaccard coefficient the bigram overlap ratio (BOR) to emphasize the meaning in our analyses. For our preliminary analyses, the set of all the texts which belong to each language is simply treated as a document, which is represented in the five ways as described above. Then, the BOR was calculated for every possible pairs of languages, and exploratory analyses were performed. Among the three normalization strategies, it turned out removal of spaces yielded poor results. For example, the results with the strategy suggested that an unexpected pair of Mizo+Uzbek was closely related, which is actually not the case: Mizo is a Tibeto-Burman language, Uzbek is a Turkic language. This pair only appears at the 45th when baseline method is employed as shown in Appendix A.

The majority of the pairs generated by the bi-gram analysis correctly grouped languages that belong to the same language family. Indo-European languages are the majority languages in our database, and many of the higher ranked pairs belong to the Indo-European languages. Other language families are underrepresented in the top 50 ranked languages in the character bi-gram analysis (Table 2).

The rankings of top 20 language pairs based on the BOR using baseline method as well as three normalizations and word bigram are shown in Table 4. After a bi-gram analysis with no diacritics in the third column, pairs except Indonesian+Malay belong to the Indo-European language family, in particular the Romance languages. The word-level bi-gram analysis also adds Bantu languages such as the Ndebele+Siswati pair or the Ndebele+Xhosa.

The bi-grams are not without their own flaw. The word-level bi-gram analysis still has pairs with Quechua and Bantu languages, which is unexpected; neither linguistically, nor geographically are these languages related. These bi-grams are thus subject to generating erroneous results. If an analyst relies on bi-gram analyses only, genetically unrelated languages could be identified as a pair of languages (such as Quechua and a Bantu language).

4.2. Hierarchical Clustering Based on Bi-gram Overlap Ratio

We conducted Ward’s hierarchical clustering to derive a tree from the BOR of every possible pairs of languages without normalization. We used the implementation of clustering available in the SciPy library\(^\text{11}^\) The result is shown in Figure 5. Germanic languages are grouped together as the green-colored cluster. The red-colored cluster comprises Romance languages. Southern Bantu languages, namely, Shona, Sesotho, Xitsonga, Tshivenda etc. are grouped in the purple-colored cluster. These results suggest that BOR metrics is reliably used for measuring similarity of languages given the parallel corpus. We also explored analyses to obtain fine-grained relationships between languages using relatively less known languages: Jinghpaw (a Tibeto-Burman language) and Ndebele (a Bantu language). In Figure 5, the top panel shows the result of hierarchical clustering using the languages of the ten highest BOR with Jinghpaw. Three languages, Angami, Mizo and Liangmai, are clustered as having closer degrees of similarity to Jinghpaw compared to other languages. If one did not have any information on Jinghpaw, now they have learned that the distance between Jinghpaw and these languages is similar. All four languages are Tibeto-Burman languages spoken in Northern Myanmar or Northeastern India. A second clustering analysis was performed with Ndebele in the same manner with the analysis of Jinghpaw. The results are in the bottom panel of Figure 5. Two languages, Siswati and Xhosa, are identified to share similar distance from Ndebele, and the results of the clustering analysis suggests that Ndebele may have more affinities with these two languages. Ndebele belongs to the same subgroup of Bantu languages called Nguni languages together with Siswati and Xhosa. What is interesting is that other southern Bantu languages (Shona, Sesotho, Xitsonga, Tshivenda, etc., non-Nguni languages) have not been identified as forming a similarity cluster as Siswati and Xhosa did. Now, the results of our preliminary analyses offer a starting point to assume that Ndebele is a Nguni language in the Bantu language family, which is indeed the case. In sum, the clustering using the BOR helps visualizing the languages that are typologically related to each other.

5. Discussion

The clustering analysis that identified related languages using a bi-gram overlap ratio is one of the many uses this database offers. About 50 languages were written in scripts that were not in the alphabetic script. This database offers a step toward developing a method comparing these scripts with others because of the

\(^{11}\text{https://docs.scipy.org/doc/scipy/reference/cluster.html}\)
Table 1: Comparing the ranked pairs according to character and word BOR; char. means character. Baseline means without any preprocessing, and the three types of preprocessing are tested with character BOR. Language pairs in red are typologically unrelated but identified as relatively similar.

| Pair                        | Baseline | char. bigram without space | char. bigram without diacritics | char. bigram lowercasing | word bigram |
|-----------------------------|----------|-----------------------------|----------------------------------|--------------------------|-------------|
| Asturian+Spanish            | Asturian+Spanish | Asturian+Spanish | Asturian+Spanish | Asturian+Spanish |
| Catalan+Spanish             | Indonesian+Malay | Catalán+Spanish | Indonesia+Malay | Catalonia+Spanish |
| Afrikaans+Dutch             | Dutch+English | Afrikaans+Dutch | Indonesia+Malay | Indonesian+Malay |
| Indonesian+Malay            | Afrikaans+Dutch | Catalán+Spanish | Catalán+Spanish | Catalán+Spanish |
| Asturian+Catalan            | Asturian+Catalan | Asturian+Catalan | Asturian+Catalan | Asturian+Catalan |
| Norwegian+Swedish           | Norwegian+Swedish | Norwegian+Swedish | Norwegian+Swedish | Norwegian+Swedish |
| Portuguese-Brazil+Spanish   | Portuguese-Brazil+Spanish | Portuguese-Brazil+Spanish | Portuguese-Brazil+Spanish | Portuguese-Brazil+Spanish |
| Catalan+Portuguese-Brazil   | Ndebele+Siswati | Portuguese-Brazil+Romanian | Catalan+Portuguese-Brazil | Catalan+Portuguese-Brazil |
| Indonesian+Mandarin         | Indonesian+Mandarin | Catalán+Portuguese-Brazil | Catalán+Portuguese-Brazil | Catalán+Portuguese-Brazil |
| K`iche`+Uspanteko           | Catalan+Portuguese-Brazil | Catalan+Portuguese-Brazil | Catalan+Portuguese-Brazil | Catalan+Portuguese-Brazil |
| Portuguese-Brazil+Romanian  | Indonesian+Acehnese | Catalan+Italian | Catalan+Italian | Catalan+Italian |
| Afrikaans+Portuguese-Brazil | Dutch+English | Asturian+Portuguese-Brazil | Asturian+Portuguese-Brazil | Asturian+Portuguese-Brazil |
| Catalan+Italian             | Dutch+English | Catalan+Italian | Dutch+English | Dutch+English |
| Catalan+English             | Dutch+English | Catalan+Romanian | Dutch+English | Dutch+English |
| Asturian+Portuguese-Brazil  | Catalan+Romanian | Catalan+Romanian | Catalan+Romanian | Catalan+Romanian |
| Dutch+U年限Deutsch          | K`iche`+Uspanteko | Indonesian+Acehnese | Catalan+French | Catalan+French |
| Dutch+Norwegian             | Dutch+Norwegian | Ndebele+Siswati | Dutch+Norwegian | Dutch+Norwegian |
| Afrikaans+Norwegian         | Dutch+Norwegian | Dutch+Norwegian | Dutch+Norwegian | Dutch+Norwegian |
| Low-Franconian+PlattDeutsch | Mizo+Uzbek | Low-Franconian+PlattDeutsch | Low-Franconian+PlattDeutsch | Low-Franconian+PlattDeutsch |

Figure 3: Clustering of languages using Latin alphabets in database based on BOR

commonalities in the text across languages; all of them contain COVID-19 related terminology and comparable contents, which facilitate such comparisons. The database has limitations as follows:

**Translation of Health related Terms** Terms such as “ultra-violet lamp”, “vaccine” were not easy to translate to some under-resourced languages because no equivalent device or concepts are in use in the languages. In most cases, the translations borrowed expressions used in the majority language nearby. While native language translations were not available, these borrowing in the translations offer anchoring points between languages that belong to different language families.

**Establishing Resource Reliability** The translations of texts were mostly done by an expert in the language. Few languages such as Nuosu Yi or Yoruba were exceptions as they were translated as a collaborative effort of specialists. Independent specialists who can evaluate the reliability of the quality of the translations need to possess expert knowledge in both the source and the target language. Most low-resourced languages do not have a pool of specialists, and 40 languages in our
Project Evaluation  The aim of this multilingual database project is to provide a multilingual clearing house for preventing the spreading of misinformation regarding COVID-19 and makes the information more accessible to as many people as possible by including low-resource languages. Evaluating the success of this kind of project requires the use of evaluative metrics, which, as far as we are aware of, is not yet available. The database can be augmented by adding user-friendly interface to add more languages, and also reflect contents that meet the needs of community members of low-resourced languages.

6. Conclusion

This paper has presented (a) a multi-lingual database created from texts of COVID-19 mythbusters from the WHO website, and (b) preliminary analyses of the these texts to figure out language family membership of unknown languages. The database includes translations of 116 versions of which 3 languages have dual orthographic conventions. All the scripts are converted to Unicode for compatibility. We have run preliminary evaluation using texts that were written in the Latin alphabet. Results of bi-gram analyses improved when diacritics were ignored in evaluating language affinity. The results of bi-gram analyses also offer insights into how parsers may succeed and fail when they compare languages that are not related. As far as the authors know, there is no comparable multilingual corpora in terms of the number of languages. We did find multiple bilingual corpora comparing a language with English such as CCAligned but no resources containing comparable texts in as large as 115 languages was found. Thus, it is currently difficult to provide the results of clustering using the BOR to other resources. When more resources such as ours become available in the future, we will be able to perform reliability tests. Evaluation using texts written in non-Latin alphabets as well as investigation on whether a tri-gram analysis or other types of text analyses would improve the evaluation will be left for future research.

7. Acknowledgements

We thank all the volunteer translators who contributed to the original language resource. The original project received support from JSPS’s Core-to-Core Program: B. Asia-Africa Science Platforms (2018-2020) “Establishment of a Research Network for Exploring the Linguistic Diversity and Linguistic Dynamism in Africa”, ICU Linguistics Lab, Mathivha Centre at University of Venda, as well as Shigeto Kawahara, P. Sreekumar, M. C. Kesva Murty, and Dr. G. Praveen. The current paper was partially supported by JSPS KAKENHI Grant (B) 21KK0005 to the 3rd author. We would like to thank the reviewers for their feedback on the manuscript.

8. Bibliographical References

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federman, C., Genzel, D., Guzm’an, F., Hu, J., Hughes, M., Koehn, P., et al. (2020). Tico19: the translation initiative for covid-19. ArXiv, abs/2007.01788.

English, P., Richardson, M., and Garzón-Galvis, C. (2018). From crowdsourcing to extreme citizen science: Participatory research for environmental health. Annual Review of Public Health, 39(1):335–350. PMID: 29608871.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online, July. Association for Computational Linguistics.

https://opus.nlpl.eu/CCAligned.php
Lee, S. J. and Won, D. (2021). Covid-19 myth busters from WHO in world languages.

Lee, S. J., Won, D., and Kawahara, S. (2021). Covid-19 myth busters in world languages: a case for broader impacts of linguistic research during the covid-19 crisis. *Reports of the Keio Institute of Cultural and Linguistic Studies*, 52:1–11, 03.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., et al. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

### A. Top 50 similar language pairs based on character BOR

| Language pair               | BOR  |
|-----------------------------|------|
| 1 Asturian+Spanish          | 0.5042 |
| 2 Catalan+Spanish           | 0.4039 |
| 3 Afrikaans+Dutch           | 0.3964 |
| 4 Indonesian+Malay          | 0.3963 |
| 5 Asturian+Catalan          | 0.3554 |
| 6 Norwegian+Swedish         | 0.3507 |
| 7 Portuguese-Brazil+Spanish | 0.3206 |
| 8 Catalan+Portuguese-Brazil | 0.3094 |
| 9 Indonesian+Mandar         | 0.3090 |
| 10 K’iche’+Uspanteko        | 0.3089 |
| 11 Portuguese-Brazil+Romanian | 0.3060 |
| 12 Indonesian+Acehnese      | 0.2984 |
| 13 Ndebele+Siswati          | 0.2941 |
| 14 Catalan+Italian          | 0.2930 |
| 15 Catalan+English          | 0.2909 |
| 16 Asturian+Portuguese-Brazil | 0.2903 |
| 17 Dutch+PlattDeutsch       | 0.2889 |
| 18 Dutch+English            | 0.2888 |
| 19 Afrikaans+Norwegian      | 0.2849 |
| 20 LowFranconian+PlattDeutsch | 0.2796 |
| 21 Catalan+French           | 0.2793 |
| 22 Italian+Spanish          | 0.2785 |
| 23 Dutch+Norwegian          | 0.2780 |
| 24 Catalan+Romanian         | 0.2778 |
| 25 Afrikaans+PlattDeutsch   | 0.2776 |
| 26 English+Spanish          | 0.2772 |
| 27 Italian+Portuguese-Brazil | 0.2689 |
| 28 Dutch+Swedish            | 0.2687 |
| 29 Afrikaans+English        | 0.2673 |
| 30 Italian+Romanian         | 0.2671 |
| 31 English+Romanian         | 0.2670 |
| 32 Afrikaans+Swedish        | 0.2648 |
| 33 Indonesian+Rejang        | 0.2642 |
| 34 Ndebele+Xhosa            | 0.2615 |
| 35 Jinghpaw+Mizo            | 0.2612 |
| 36 Dutch+LowFranconian      | 0.2582 |
| 37 Asturian+English         | 0.2573 |
| 38 English+Portuguese-Brazil | 0.2559 |
| 39 English+Norwegian        | 0.2553 |
| 40 Liangmai+Jinghpaw        | 0.2547 |
| 41 French+Spanish           | 0.2542 |
| 42 English+French           | 0.2542 |
| 43 Romanian+Spanish         | 0.2526 |
| 44 Asturian+Italian         | 0.2518 |
| 45 Mizo+Uzbek               | 0.2513 |
| 46 French+Italian           | 0.2506 |
| 47 Mandarin+Acehnese        | 0.2503 |
| 48 Turkish+Uzbek            | 0.2499 |
| 49 Malay+Acehnese           | 0.2497 |
| 50 Asturian+Romanian        | 0.2483 |

Table 2: Top 50 similar language pairs based on character BOR. Language pairs in red are typologically unrelated but identified as relatively similar.
Table 3: Descriptive statistics of 65 languages that use Latin alphabets in the dataset. The column Ave. words indicates the average number of words per one myth buster text in a language. The last column lists five-most frequent words with more than four characters when lowercased.