Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3048–3055
Marseille, 20-25 June 2022

COVID-19 Mythbusters in World Languages

Mana Ashida[1], **Jin-Dong Kim**[2], **Seunghun J. Lee**[3,4]
[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 languages[1]. 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 languages[3]. Endangered languages Fund has a resource website for languages that are indigenous, endangered or under-resourced[4].A 5-phrase translation

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[2] https://www.covid19parenting.com/#/tips

[3] https://translatorswithoutborders.org/translations-covid-19/

[4] https://endangeredlanguagesproject.github.io/COVID-19/

project for hygiene awareness used an easy-to-input interface so that the 5 phrases become available in as many languages as possible.

TICO-19 (Anastasopoulos et al., 2020) is a similar project on terminologies concerning COVID-19 and public heath. The Asian & Pacific Islander American Health Forum offers a webpage with COVID-19 related terms and phrases in 5 Pacific languages[5].

These multilingual projects about the COVID-19 pandemic had a goal of spreading important information to many language speakers during the onset of the pandemic, which is an aim that is comparable to the Mythbusters project. However, other projects have not yet made their text resources available in a single database that can be used by the community of NLP researchers. The COVID-19 mythbusters multilingual database provides an opportunity for cross-linguistic data analyses.

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[6]", 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.

[5] https://www.aa-nhpihealthresponse.org/nh-pi-translated-covid-19-terms

[6] https://www.masakhane.io/

3. Project Overview

3.1. Mythbusters in World Languages

The main website[7] 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"[8].

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. Disclaimers were part of all the translations because they were not created by the WHO. Pages of 13 languages also include a link to a Youtube video where the texts are narrated by a native speaker[9]. One language has audio files only of this narration. Languages with the audio files often have speakers who have low literacy in their own script, or where speakers are not used to see their own language in a text format. Detailed descriptions of the creation of this website can be found in Lee et al. (2021).

The text of the mythbusters were processed and cleaned up by the first author, and database structure was created by the second author. The collaboration of the three authors resulted in a web database that contained freely downloadable texts of COVID-19 mythbusters in more than 100 languages. This database is unique due to two points. First, while the amount of texts in each language is relative small, this database is one of the few that consists of a diverse set of languages whose texts are directly comparable. Second, most of the translations were created by native speaker professionals who specialize in their language or who work on translations to and from their language.

3.2. Structure of the Database

In January 2022, the database contains text of 116 versions that is accessible from `https://db.covid-no-mb.org`; ISLRN 620-293-539-189-9. 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

[7] https://covid-no-mb.org/

[8] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

[9] https://www.youtube.com/channel/UC2zlmp0_cp3DUzjPrnT5jPw

Figure 1: Homepage of the Database

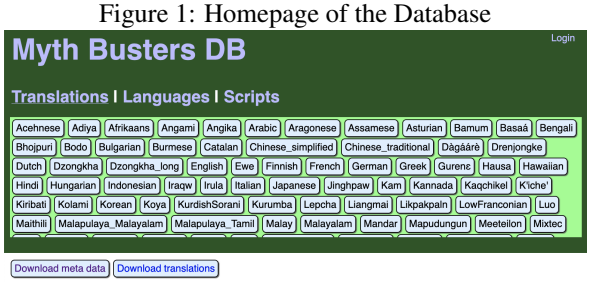
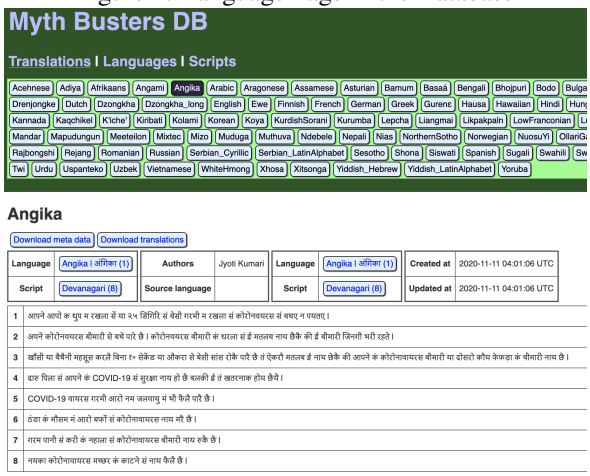


Figure 2: Language Page in the Database



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, Hangeul, 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[10], 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.

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

Hereafter, we call the Jaccard coefficient the *bigram overlap ratio (BOR)* to emphasize the meaning in our implementation. 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 1. After a bi-gram analysis with no diacritics in the third column, pairs except Indonessian+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-gram analyses 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-gram analyses 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[11]. The result is shown in Figure 3. 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 4 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

[10] acute, grave, umlaut, trema etc.

[11] `https://docs.scipy.org/doc/scipy/reference/cluster.html`

	baseline	char. bigram without space	char. bigram without diacritics	char. bigram lowercasing	word bigram
1	Asturian+Spanish	Asturian+Spanish	Asturian+Spanish	Asturian+Spanish	Asturian+Spanish
2	Catalan+Spanish	Indonesian+Malay	Catalan+Spanish	Afrikaans+Dutch	Indonesian+Malay
3	Afrikaans+Dutch	Catalan+Spanish	Afrikaans+Dutch	Catalan+Spanish	Catalan+Spanish
4	Indonesian+Malay	Afrikaans+Dutch	Indonesian+Malay	Indonesian+Malay	Indonesian+Mandar
5	Asturian+Catalan	Asturian+Catalan	Asturian+Catalan	Asturian+Catalan	Ndebele+Siswati
6	Norwegian+Swedish	Norwegian+Swedish	Norwegian+Swedish	Norwegian+Swedish	Norwegian+Swedish
7	Portuguese-Brazil+Spanish	Portuguese-Brazil+Spanish	Portuguese-Brazil+Spanish	Portuguese-Brazil+Spanish	Asturian+Catalan
8	Catalan+Portuguese-Brazil	Ndebele+Siswati	Portuguese-Brazil+Romanian	Catalan+Portuguese-Brazil	Ndebele+Xhosa
9	Indonesian+Mandar	Indonesian+Mandar	Catalan+Portuguese-Brazil	Catalan+Italian	Quechua+Xhosa
10	K'iche'+Uspanteko	Catalan+Portuguese-Brazil	Asturian+Portuguese-Brazil	Indonesian+Mandar	Shona+Siswati
11	Portuguese-Brazil+Romanian	Indonesian+Acehnese	Catalan+Italian	K'iche'+Uspanteko	Ndebele+Quechua
12	Indonesian+Acehnese	Portuguese-Brazil+Romanian	Indonesian+Mandar	Asturian+Portuguese-Brazil	Quechua+Uzbek
13	Ndebele+Siswati	Asturian+Portuguese-Brazil	Catalan+French	Catalan+Romanian	Dutch+Quechua
14	Catalan+Italian	Dutch+English	Catalan+English	Indonesian+Acehnese	Finnish+Xhosa
15	Catalan+English	Catalan+English	Catalan+Romanian	Portuguese-Brazil+Romanian	Finnish+Quechua
16	Asturian+Portuguese-Brazil	Catalan+Romanian	K'iche'+Uspanteko	Catalan+French	Siswati+Xhosa
17	Dutch+PlattDeutsch	K'iche'+Uspanteko	Indonesian+Acehnese	Catalan+English	Quechua+Siswati
18	Dutch+English	Catalan+Italian	Ndebele+Siswati	Dutch+English	Shona+Xhosa
19	Afrikaans+Norwegian	Dutch+Norwegian	Dutch+PlattDeutsch	Italian+Spanish	Finnish+Ndebele
20	LowFranconian+PlattDeutsch	Mizo+Uzbek	Afrikaans+Norwegian	Dutch+PlattDeutsch	Quechua+Shona

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.

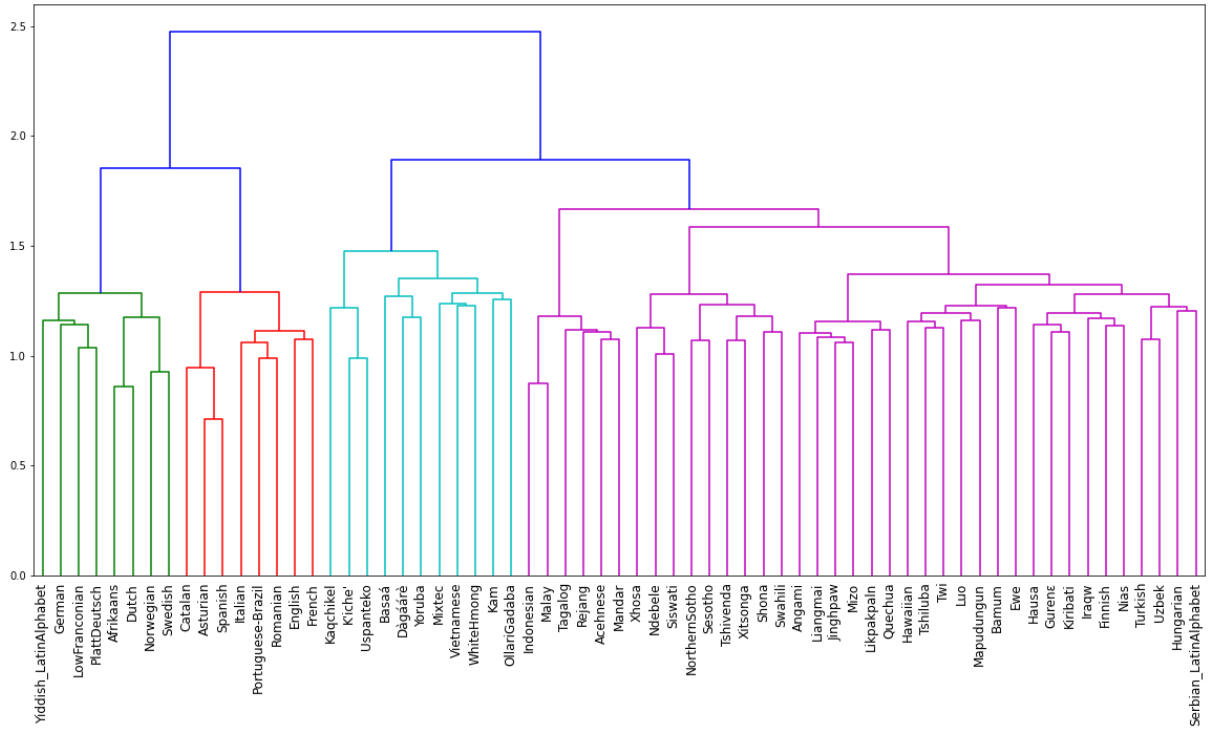


Figure 3: Clustering of languages using Latin alphabets in database based on BOR

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

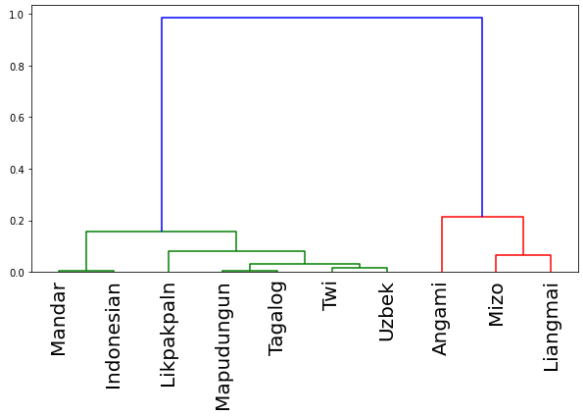


Figure 4: Clustering using the top ten closest languages to Jinghpaw

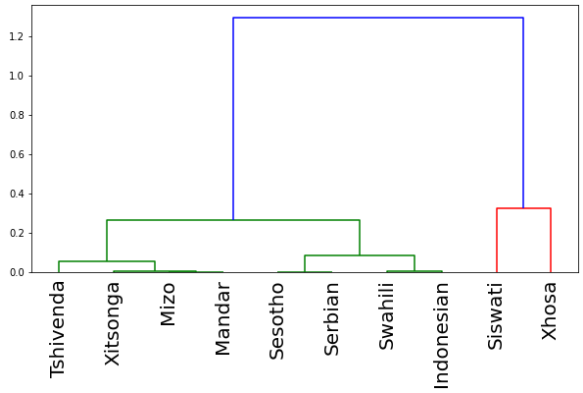


Figure 5: Clustering using the top ten closest languages to Ndebele

database do not have any pages on wikipedia. We consider this database as a starting point for building multilingual resources that include and also benefit low resourced languages.

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 [12], 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). Tico-19: 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.

[12] 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	Mandar+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.

B. Average length and word-frequency of mythbusters in the database

language	Ave. words	Top-5 most frequent words (word len. more than 4 characters)
Acehnese	20.74	corona (20), taeun (19), baroe (15), atawa (12), hanjeut (9), penyaket (8)
Afrikaans	19.54	koronavirus (15), voorkom (7), covid-19 (7), mense (6), ander (5), koronavirussiekte (4)
Angami	14.25	coronavirus (20), morei (14), kechüu (10), kechü (7), geinu (7), kenjü (7)
Asturian	17.38	coronavirus (22), (covid-19) (20), nuevu (9), prevenir (4), previén (3), enfermedá (3)
Bamum	27.84	coronavirus (25), mànjé (9), (covid-19) (5), yétne (5), nshié (5), yetne (4)
Basaá	19.26	corona (21), ńsɔ́ŋ (15), bí!sú (4), mbóóp (3), lɛ́ɛ́gɛ́ (3), ń!lék (2)
Catalan	18.75	coronavirus (16), covid-19 (9), prevé (4), prevenir (4), (covid-19) (3), malaltia (3)
Dàgáárè	16.46	corona (26), bààlɔ́ (25), bíí (11), wúlì (6), pɔ́ɔ́(5), zàá (5)
Dutch	18.89	coronavirus (19), nieuwe (17), mensen (6), tegen (5), worden (5), voorkomt (4)
English	18.5	coronavirus (13), prevent (8), covid-19 (8), corona (6), people (6), disease (5)
Ewe	23.0	koɖonavaɖɔsi (17), dɔlélе (9), dɔlélea (8), yeyea (6), dɔlékui (5), aɖeke (4)
Finnish	11.21	koronavirusta (7), eivät (6), uutta (6), koronavirus (3), (covid-19:ää) (2), koronaviruksesta (2)
French	21.25	coronavirus (15), nouveau (14), covid-19 (8), maladie (5), peuvent (5), votre (5)
German	19.84	coronavirus (18), neuen (13), nicht (11), einer (8), ansteckung (4), werden, (4)
Gurenɛ	25.68	korona (20), ta'am (19), varusi (19), nyɔkɛ(7), tulegɛ(5), magesiini (4)
Hausa	19.5	cutar (23), covid-19 (11), kwaronabairas (9), sabuwar (8), kamuwa (6), kariya (4)
Hawaiian	29.33	coronavirus (18), covid-19 (8), kekahi (5), kaohi (4), kokua (4), lapaau (4)
Hungarian	17.53	koronavírusos (9), (covid-19) (5), fertőzéstől (4), koronavírust (4), valaki (3), pusztítja (3)
Indonesian	18.79	virus (21), tidak (16), corona (14), dapat (11), mencegah (7), penyakit (6)
Iraqw	16.42	koroona (20), firuusír (18), /aben (15), bará (11), laqaá (8), covid- (7)
Italian	17.17	covid-19 (12), coronavirus (11), nuovo (11), prevenire (8), aiuta (6), essere (6)
Jinghpaw	16.89	coronavirus (21), kahtet (6), shing (3), makawp (3), lahta (2), machyi (2)
K'iche'	20.83	yab'il (23), xuquje' (17), k'ak' (12), coronavirus (11), kuq'atej (8), (2019-ncov) (8)
Kam	19.25	bingh (17), yongh (9), mangc (8), gueec (8), naengl (7), fangc (6)
Kaqchikel	20.83	rik'in (17), coronavirus (17), k'ak'a' (13), richin (12), covid-19 (8), (2019-ncov) (8)
Kiribati	30.33	aoraki (20), coronavirus (18), kabuebue (7), reken (6), ibukin (5), aomata (5)
Liangmai	16.12	corona (28), tiubo (7), chapiu (5), wikhaibo (5), kamsat (4), marabo (3)
Likpakpaln	21.24	corona (23), virus (22), aaween (22), machine (5), ninchee (4), chuur (4)
LowFranconian	16.46	corona-virus (10), covid-19 (10), corona-seeke (7), (covid-19) (4), kriegen (4), hölpt (3)
Luo	14.92	korona (24), geng'o (4), nyaka (4), gengo (4), kong'o (3), ariyo (2)
Malay	17.37	tidak (16), boleh (12), covid-19 (10), coronavirus (10), jangkitan (8), penyakit (6)
Mandar	13.25	virus (18), corona (14), andiangi (12), covid-19 (7), andiang (5), mipakarao (5)
Mapudungun	16.53	kutran (22), koronafirus (14), (covid-19) (9), firus (5), korona (4), pülku (3)
Mixtec	19.71	kuè'è (25), korónavírus (23), xí'in (9), ki̱'in (7), tàtán (6), (covid-19) (5)
Mizo	25.37	coronavirus (21), natna (8), lakah (6), theih (5), zawng (4), theihna (4)
Ndebele	13.11	nanyana (9), ye-coronavirus (9), ingogwani (7), abantu (6), i-coronavirus (5), ngabe (4)
Nias	19.5	virus (20), korona (17), igöna (13), khöda (9), fa'atola (8), covid-19 (8)
NorthernSotho	29.5	corona (20), baerasi (17), bolwetši (16), bofsa (14), batho (7), twatši (6)
Norwegian	18.42	koronaviruset (16), eller (10), smitte (5), smittet (4), drepe (4), koronaviruset? (4)
OllariGadaba	18.22	karana: (14), pa:iṭ (11), niya: (10), kegiŋ (9), rakhya: (7), ering (6)
PlattDeutsch	17.21	koronavirus (10), covid-19 (10), virus (8), töggen (7), vehinnern (6), vespreden (4)
Portuguese-Brazil	19.84	coronavírus (14), causadas (6), pessoas (6), estão (5), doenças (4), prevenir (4)
Quechua	11.58	manan (14), musuq (12), (covid-19) (11), coronavirus (8), coronavirusmanta (5), amachakushanchischu (2)
Rejang	17.29	korona (20), virus (18), têmgêak (9), kundêi (8), panês (7), infeksi (6)
Romanian	19.58	coronavirus (13), poate (6), covid-19 (4), împotriva (4), eficiente (4), noului (4)
Serbian	15.38	novim (11), koronavirusom (11), kovid-19 (5), sprečava (4), koronavirus (4), osobe (4)
Sesotho	29.44	kokwana-hloko (23), corona (17), (covid-19) (7), tshwaetso (5), thibela (5), karabo: (5)
Shona	16.68	chekoronavhairasi (14), chirwere (13), vanhu (6), (covid-19) (5), kubva (5), kuchirwere (5)
Siswati	18.06	lekhorona (17), leligciwane (17), lelivela (9), kamuva (6), (covid19) (6), bantfu (6)
Spanish	18.83	coronavirus (16), nuevo (13), covid-19 (8), (2019-ncov) (8), puede (6), previene (4)
Swahili	21.89	virusi (22), korona (18), vipya (17), ugonjwa (5), katika (5), kukinga (4)
Swedish	16.68	coronaviruset (13), eller (9), effektiva (4), förhindra (4), personer (4), coronavirussjukdomen (3)
Tagalog	17.42	hindi (22), covid-19 (14), coronavirus (9), maaagapan (5), virus (5), kahit (4)
Tshiluba	17.79	coronavirus (25), disama (20), mubidi (8), buanga (7), mudisama (6), nshebeya (5)
Tshivenda	23.5	khorona (24), tshitzhili (23), vhathu (6), thivhela (5), kavhiwa (4), thivheli (3)
Turkish	13.96	koronavirüse (10), koronavirüs (8), karşı (8), tedavi (5), sıcak (3), engellemez (3)
Twi	22.26	coronavirus (14), ntumi (9), covid-19 (5), entumi (5), biara (4), nnuru (4)
Uspanteko	19.79	yajeel (27), suteem (26), kita' (20), (2019-ncov) (7), looq' (6), kita' (6)
Uzbek	15.79	yangi (17), koronavirus (10), koronavirusni (8), olish (6), oldini (5), sizni (4)
Vietnamese	19.78	không (19), virus (19), corona (19), nhiễm (9), những (4), chống (4)
WhiteHmong	31.75	covid-19 (21), thaiv (11), thiab (9), txawm (8), tseem (6), tshuaj (6)
Xhosa	8.83	icovid-19 (14), okanye (7), ukuba (3), kwi-covid-19 (3), kwaye (3), inganobungozi (3)
Xitsonga	29.39	xitsongwatsongwana (21), khorona (21), lexintshwa (17), kumbe (12), vanhu (7), vuvabyi (5)
Yiddish	19.32	korone-virus (18), nisht (17), (covid-19) (9), nayem (8), kenen (7), onshtekn (4)
Yoruba	24.16	kòkòrò (19), kòrónà (19), àrún (16), titun (15), tàbí (11), àwọn (8)

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.