Developing an Up-to-date Academic Word List for Public Health Emergencies of International Concern: The Case of Mpox

Longxing Li

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Developing an Up-to-date Academic Word List for Public Health Emergencies of International Concern: The Case of Mpox

Longxing Li
Faculty of Languages and Translation
Macao Polytechnic University
Rua de Luís Gonzaga Gomes, Macao, China
lxli@mpu.edu.mo

Abstract

The mpox epidemic in many countries has made it a Public Health Emergency of International Concern for two times so far. Learning lessons from the COVID-19 pandemic, the international community has made response plans and proposed measures to contain the epidemic. From the perspective of emergency language service, this paper aims to develop an academic word list on mpox to assist prevention and facilitate communication. The powerful corpus tool Sketch Engine and its rich corpus resources are exploited to make the Mpox Word List, which includes about 300 words. The four-step purpose-oriented procedure for academic word list development is synthesized based on previous literature and the practice of developing this mpox word list. Further research directions in extracting multi-word terms or n-grams and developing supplementary acronym list can be included to facilitate academic exchange and mass communication. This paper is believed to be of significant reference value for further term extraction practice and word list development in public emergencies, crises, and other scenarios.

1 Introduction

In 1970, the first human cases of mpox were diagnosed in the Democratic Republic of the Congo. Decades after, a global outbreak of clade II mpox in 2022–2023 marked the first incidence of widespread community transmission outside of Africa. In July 2022, the World Health Organization (WHO) declared the outbreak a Public Health Emergency of International Concern (PHEIC) which means a public health risk requiring immediate international action. The declarations of PHEIC can rapidly mobilize international coordination, streamline funding, and accelerate the advancement of the development of vaccines, therapeutics and diagnostics. The ultimate purpose of such declaration is to catalyze timely worldwide and evidence-based action to limit the societal impacts of emerging and re-emerging disease risks (Wilder-Smith & Osman, 2020). In May 2023, as the outbreak of mpox came under control, the PHEIC status was reverted. However, a later outbreak of clade I mpox detected in the Democratic Republic of the Congo during 2023 has spread to several African countries. On 14 August 2024, the WHO declared this outbreak a PHEIC again. As of 16 August 2024, fifteen countries were reported to have identified cases of mpox, with 16,839 reported cases and 501 reported deaths (a fatality rate of about 3-4%).

To respond to this PHEIC, the WHO issued Global Strategic Preparedness and Response Plan and proposed mpox control and elimination plans in the Strategic Framework for Enhancing Prevention and Control of mpox 2024–2027. This framework emphasizes integrating efforts of all health programs and coordination among all partners and stakeholders to ensure a continued and robust response for mpox. International risk communication, community engagement, and patient services are essential components to control mpox outbreaks in every context. The clear and consistent terminology is a requisite for quick and smooth communication. Terminology support provided in the forms of academic word lists and terminology management play an important role in public risk communication and emergency language service.

The COVID-19 pandemic in the past few years made us acutely aware of the importance of language services in international public health emergencies. To facilitate communication during the prevention and control of the pandemic, many researchers have developed various word lists with specific purposes and functions. Li and Wang (2021) extracted 364 single-word and 176 multi-word COVID-19 terms using the COVID-19 thematic academic corpus and the corpus tool Sketch Engine. Saed et al. (2022) established a COVID-19 lemmatized word list classified into six categories using Factiva news data and the corpus tool Wordsmith.

At present, there are few attempts at producing a mpox word list. Although COVID-19 and mpox are not spreading on the same scale, there is still a need to prepare for a mpox epidemic on a larger scale. Therefore, learning from previous researchers' experience in producing academic word lists, especially COVID-19 word lists, this paper attempts to develop a specific mpox word list to help its prevention and control.

The attempts of making word list for specific purposes can be traced back to the 1950s. To meet the needs in teaching and learning English vocabulary, West (1953) developed one of the earliest word lists, the General Service List (GSL), in which 2,000 common word families were listed. In recent decades, researchers have expanded the field of word list development by making lists for more specific purposes. The difficulty and importance of medical vocabulary make the word list of English for medical purposes (EMP) an important branch, e.g., the Medical Academic Word List (MAWL) by Wang et al. (2008) and the Medical Academic Vocabulary List (MAVL) by Lei and Liu (2016). Different from MAWL, MAVL only includes words with the minimum frequency higher than 28.57 times per million words (PMWs) to ensure that the included words are frequently used. In developing the COVID-19 word list, Li and Wang (2021) addressed existing problems in previous studies and proposed a purpose-oriented five-step procedure for word list development. The standardized procedure and the way Saed et al. (2022) categorize the words in the list will be referenced to guide the development of the Mpox Word List in the current study.

2 Methods

2.1 The corpus tool Sketch Engine

Sketch Engine (Kilgarriff et al. 2004, Kilgarriff et al. 2014) is the representative of the fourth-generation corpus retrieval tools, which realizes the online retrieval of corpus and offers the core functions: word sketch, word sketch difference, thesaurus, concordance, wordlist, keywords, and n-grams. It has been widely used in lexicology, language teaching, discourse analysis, translation studies, contrastive linguistics, keyword studies, and so on as in Li et al. (2018), Li et al. (2020), Li et al. (2021), and Li (2023). The keywords function is the most important in the production of word lists.

The keywords function of Sketch Engine is designed to compare two corpora to find out the unique or typical words in one corpus relative to the other corpus; these words can help understand the contents or topics of the corpus, so this function is especially suitable for retrieving keywords or extracting terms. The selection of the reference corpus determines the relevance of the extracted candidate items to the topic. Taking the production of the COVID-19 Word List as an example, the word list with EGP corpus as a reference corpus may contain a large number of general medical expressions and has a weaker correlation with the topic than that with EMP corpus as a reference. However, if another medical corpus in Sketch Engine is adopted as a reference corpus, the pertinence and emergency of the COVID-19 Word List can be improved and a large number of general medical words possibly known to the word list users can be reduced. The size of the focus corpus used for terminology extraction does not need to be very big, but a larger corpus covers more terms. The larger the reference corpus is, the better. The size of the COVID-19 corpus and other medical corpora in Sketch Engine is large enough to meet the requirements for the production of the word list.

Keywords are single-word items that appear more frequently in focus corpus than in reference corpus, which can be displayed in lemma, word, or other forms according to the needs and are case sensitive. In other words, the keywords function can select the displayed form of words according to the needs of researchers and extract the single-word terms and multi-word terms simultaneously, thus it significantly enhances the efficiency of word list development. From the introduction above, it can be seen that Sketch Engine is a corpus tool

suitable for providing emergency terminology services.

2.2 Corpus data

The available academic and language data for mpox is not so rich as those for COVID-19 which has a 1.4-billion-word specialized medical corpus CORD-19 in Sketch Engine. So, it would be more challenging to select proper data resources for developing the Mpox Word List. To collect a dataset large enough for keywords extraction, the rich corpora resources in Sketch Engine are explored and the corpus English Trends (2014–today) is considered as a possible source of data. The English Trends Corpus (ETC hereafter) is a monitor corpus consisting of news, research articles, Wikipedia, and texts from other sources. The corpus has been regularly updated with new texts since 2014 and grows by about 70 million words every week. As of August 2024, the corpus has reached a size larger than 82 billion words. The considerable size and its timely update make it a valuable resource to generate sufficient data for emerging or less-frequently discussed topics. A good amount of concordances of monkeypox and mpox are expected to be produced from the ETC, thus a concordance corpus of a reasonable size with mpox as its theme can be built to extract terms.

The next step is to decide which word to be retrieved as the keyword (KWIC) to create the concordance corpus. It should be pointed out that the name of the disease has been changed from monkeypox to mpox now. On November 28, 2022, WHO announced that Mpox should be used to refer to the disease and monkeypox would phase out in one year to reduce stigma, discrimination, and racism against certain animals and groups of people (Damaso, 2022). Therefore, to elicit a wider coverage of data on the epidemic, both words should be retrieved to generate more relevant terms. On 25th August 2024, the author retrieved *monkeypox* and *mpox* in the ETC by confining the data within the academic, encyclopedia, and news genres. The retrieval produced 93,890 and 4,444 occurrences of *monkeypox* and *mpox* respectively, which is large enough for building a concordance corpus on mpox. Due to the limit set by Sketch Engine, the maximum number of lines downloadable for a retrieved concordance is 10,000 with each line having a context of 100 characters left and right of the KWIC. So, 10,000 randomized concordance lines of *monkeypox* and all the 4,444 concordance lines of *mpox* are downloaded to create the Mpox Corpus. The corpus has a total of 671,856 tokens or 553,419 words. Considering that the concordances retrieved from the news genre are included in the Mpox Corpus, another concordance corpus named Mpox Academic Corpus with texts from only the academic and encyclopedia genres is created. The Mpox Academic Corpus, made of 1,244 concordances of monkeypox and 433 concordances of mpox, has a much smaller size of 75,931 tokens or 59,442 words. Later both corpora will be compared as focus corpora in terms of the effectiveness in extracting terms and the relevance of terms extracted.

To extract terms, a reference corpus should be selected to compare with the focus corpus. In Sketch Engine, by default, the largest corpus in the language is selected as the reference corpus to represent general language. This setting tends to generate a longer list of basic terms for academic or medical purposes. To make the list more relevant to mpox, an academic subcorpus of general medical and biological sciences is created from the Directory of Open Access Journals (DOAJ) corpus in Sketch Engine. DOAJ is composed of papers published in open-access journals in all areas of science, technology, medicine, social sciences, and humanities. The corpus is large with 2.6 billion English words and up-to-date with about 90% of the texts published between 2000 and 2017. The rich metadata such as the journal title, country, year of publication, publisher, subject, and article title are retained to facilitate the creation of subcorpora according to different needs. A subcorpus is created using the data from 2014 to 2017 under the subjects related to biological, medical, and health sciences. The subcorpus reaches a size of 8,220,236 words and is named DOAJ-BioMed.

The two mpox corpora and the DOAJ-BioMed corpus are thus diachronically and thematically comparable on the same platform and will be used in producing the word lists. Part 3 will introduce in greater detail the application of Sketch Engine in the production of mpox single-word list.

3 Producing the Mpox Word List

Inspired by the five-step purpose-oriented procedure in word list production by Li and Wang (2021), the author synthesized the procedure into four:

(1) analyze the purpose and users' needs of the word list and formulate principles of the word list development accordingly;
(2) set quantitative and qualitative criteria for item screening based on the principles;
(3) improve the word list by consulting users and medical professionals before publishing the list; and
(4) publish and update the word list.

3.1 Needs analysis and principles for Mpox Word List production

As previously mentioned, the word list serves medical workers, researchers, teachers and students, journalists, and common people, to meet their needs of academic activities, teaching and learning, publicizing, reporting and any other forms of communication. Most users of the word list are expected to be professionals with certain medical knowledge or with a higher education background. Therefore, two basic principles are formulated for producing the word list: first, only terms frequently appearing in the mpox research are included; second, the included terms should also be highly relevant to mpox research. The two principles ensure the high frequency and relevance of the terms listed, thus easing the burden on users and students.

3.2 Terms retrieval and the quantitative selection criteria

Guided by the two principles, the author decided on a specific term retrieval plan and the automatic selection criteria. The Mpox Academic Corpus and the Mpox Corpus are the two potential focus corpora to extract terms. There will be a comparison between the rate of terms accepted from the automatically generated keywords list. The DOAJ-BioMed subcorpus will be the reference corpus.

When retrieving the candidate terms using Sketch Engine's keywords function, most default settings are kept unchanged. For example, the "focus on" value is kept at "1" to elicit words rarely or seldomly used in general language or the reference corpus, which is more suitable for terminology extraction. The option "at least one alphanumeric" is ticked to include the retrieved lexical phrases containing at least one letter or number, such as *10-year-old* and *5G*. The maximum number of candidate terms is set to 1,000, and the single-word items are displayed in lemma (See Figure 1).

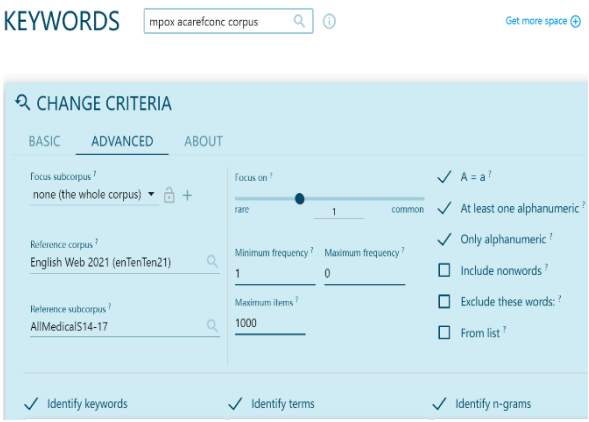

Figure 1: Keywords retrieval interface and settings.

The lists were produced and the retrieval results were saved as Excel files. Figure 2 shows the top 10 candidate single-word terms ranked by the keyness score extracted from the Mpox Academic Corpus. Two rounds of selection are conducted after the automatic generation of candidate terms. The first round of selection of terms is based on the relative frequency and keyness score of the retrieved items. Single-word terms and MWEs may follow different criteria in relative frequency and keyness score. Based on the review of the criteria (Li and Wang, 2021) in previous word list development practice, the minimum relative frequency for the included single-word terms is set at 30 PMWs.

	Lemma	Frequency per million Focus	Reference	Score
1	monkeypox	22,362.41	0.00	22,363.4
2	mpox	7,928.25	0.00	7,929.3
3	mpxv	1,962.31	0.00	1,963.3
4	smallpox	2,449.59	0.29	1,899.6
5	covid-19	1,185.29	0.00	1,186.3
6	tecovirimat	553.13	0.00	554.1
7	orthopoxvirus	487.28	0.00	488.3
8	gbmsm	474.11	0.00	475.1
9	pages	737.51	0.58	467.4
10	drc	737.51	0.68	440.4

Figure 2: Top ten single-word terms ranked by keyness score.

The keyness score is a value used by Sketch Engine to determine the particularity of a certain item in the focus corpus relative to the reference corpus. The higher the keyness score is, the more

prominent the word will be in the focus corpus. Therefore, it reflects the characteristics of the focus corpus and the possibility of a word to be selected as a term. However, there are few studies on the selection criteria based on keyness scores. The score should be varied in different cases. We can determine the score by considering the purpose and appropriate size of the word list. In this paper, the threshold for including single-word terms is set as: keyness score > 25. There are 327 candidate single-word items that meet the two criteria.

3.3 Manual selection of terms

The second round is manual selection based on the author's experience and expertise and consultation with the original context of the terms. Errors, general words, and those less relevant to mpox or the public emergency such as *say*, *pages*, *nation*, and *hong*, are excluded from the candidate list. The shortened list enables the users of the word list to focus on the terms highly related to the topic and improves communication efficiency during the public health emergency.

The manual selection excluded about 10% of the candidate terms extracted from the more specialized Mpox Academic Corpus, which is lower than that from the Mpox Corpus which includes news texts. Therefore, the list produced from the Mpox Academic Corpus, which includes 277 terms, is adopted as the Mpox Word List. This further illustrate that the size of the corpus may not be a barrier for term extraction as long as the focus corpus is properly constructed with a prominent theme.

3.4 The Mpox Word List

Due to the limited space, part of the selected terms ordered by their relative frequency in the focus corpus Mpox Academic Corpus are listed in Table1. The list is presented simply in two grades with the 100-PMWs frequency as the divide. The categorization is flexible subject to users' needs and specific contexts.

Grade I (frequency≥100 PMWs) monkeypox, virus, mpox, vaccine, outbreak, africa, spread, smallpox, mpxv, vaccination, united, acceptance, covid-19, scientist, intention, vaccinate, drc, kingdom, nigeria, zoonotic, tecovirimat, public-health, pandemic, congo, orthopoxvirus, gbmsm, variola, vaccinia, credit, endemic, democratic, sars-cov-2, cdc, warn, worry, non-endemic, plhiv, fluid-filled, mva-bn, lgbtqi, cowpox, a29, gay, vacv, poxviruse, epidemiologist, sub-saharan, orthopoxviruse, announce, med, rimoin, semen, briefing, multi-country, guangdong, portugal, ebola, pheic, mccollum, bodily, lewis, yinka-ogunleye, stockpile, virologist, msm, campus, a27, getty, mpx, bisexual, curb, wealthy, alarm, eradicate, zaire, varv, afp, coronavirus, ogoina, poxviridae, camelpox, seifert, pso, egger, lancet, virological, fatality, infectious-disease, orthopox, non-gbmsm, squirrel, begg, importation, reversion, cholera, archived, unrecognized, prep, surge, post-exposure, ukhsa, grapple, poxvirus, neglected, cynomolgus, containment, deadly
Grade II (frequency＜100 PMWs) lefkowitz, a29-specific, rosamund, rabbitpox, sars-cov-1, medrxiv, hesitancy, amr, ankara, human-to-human, mystery, flu-like, prairie, coloured, pod, transmitted, zika, tame, mers-cov, jynneos, monkeypox-related, director-general, wpr, imvanex, cpxv, non-binary, icalmed, non-african, chunk, hooper, macintyre, angeles, confirmed, transgender, wane, acing, announcement, atlanta, nordic, nat, universities, dnas, authorize, self-sampling, acam2000, cceptance-uptake, transm, mousepox, dimie, cmlv, mbala, lethally, spill, plasmablast, scab, traveler, adept, medium-term, high-income, zoonosis, vigilance, chickenpox, prophylactically, emerging, quarantine, measles, assault, georgia, zoo, rename, pury, neuropsychiatrist, anti-vacv, sometimes-painful, population-wide, heavy-tailed, abuja, orthopoxviruses, virol, yola, hatcher, seminary, mpxvs, conspiracy, hics, covid, sigh, heed, eess, heterotypic, unheeded, trialling, twitter, microevolution, siga, Liberia, skin-to-skin, sars, soar, laboratory-confirmed, wake-up, pledge, reg, scourge, bioterrorism, re-evaluate, asymptomatically, unnoticed, spark, generalist, genre, Utrecht, rope, pep, weakened, towel, peruvian, unvaccinated, stark, re-evaluated, msld, adesola, mpvx, nonendemic, basankusu, tpoxx, eurosurveillance, malembaka, ferré, happi, smallpox-like, pepv, worried, cnns, anti-mpxv, accuse, mass-vaccination, cd3-cd19, anteater, optimization-based, phylogenomic, inbox, uptake, vaccination, linelist, mpox-related, cuimc, convolutional, ntc, ectromelia, pre-outbreak, two-dose, swed, ncdc, mononucleosis-like, glimmer, reemergence, jab, computer-aided, Haiti, s2b, explode, dean, sentiment, treaty, leave-one-out, gambian, attendee, coronavirus, polio,…

Table 1: The selected terms in the Mpox Word List

4 Conclusion

The development of word lists is a fundamental task and a prerequisite for many other emergency services, such as standardization of terminology, emergency medical interpreting and translation,

construction of terminology translation database, machine translation, academic vocabulary teaching and learning, and mass communication. Aiming at providing terminology support as part of the emergency language service for mpox prevention and control, the author clarified the needs for and the purpose of producing the Mpox Word List. Following the principles and criteria in extracting and including terms, the Mpox Word List has been efficiently produced using the corpus tool Sketch Engine and its rich medical corpus resources. The synthesized purpose-oriented procedure for word list development can be used to guide subsequent development of word lists for other specific purposes.

In the future, the multi-word terms or n-grams can be included in the list to cover more essential terms for academic exchange and mass communication. A supplementary acronym list with the full spelling and definitions or explanations of these acronyms can also be made when there are a big number of them to meet the needs of the public, especially the non-professional users. The easy access to the original context of the KWIC and the retrieved items and the embedded Wikipedia links for them can be of tremendous help in developing such lists. In addition, cooperation with users and professionals from various disciplines in the development and application of the word lists should be strengthened, and feedback from medical experts and users should be collected to improve and update the word list on a regular basis. The word list produced in this study will help providers and consumers of emergency language services during international public health emergencies and the method and practice introduced in this paper is also believed to benefit future academic language researchers and academic word list developers.

Acknowledgments

The author would like to acknowledge the three anonymous reviewers of the paper and the conference participants for their valuable comments and feedback.

References

Annelies Wilder-Smith, and Sarah Osman. 2020. Public health emergencies of international concern: a historic overview. *Journal of Travel Medicine*, 27(8), taaa227. https://doi.org/10.1093/jtm/taaa227

Clarissa R. Damaso. 2023. Phasing out monkeypox: mpox is the new name for an old disease. *The Lancet Regional Health–Americas*, 17: 100424. https://doi.org/10.1016/j.lana.2022.100424

Jing Wang, Shao-lan Liang, and Guang-chun Ge. 2008. Establishment of a Medical Academic Word List [J]. English for Specific Purposes, (4): 442–458. https://doi.org/10.1016/j.esp.2008.05.003

Kilgarriff A, Rychlý P, Smrz P, et al. 2004. The Sketch Engine. Proceedings of the Eleventh EURALEX International Congress.

Kilgarriff A, Baisa V, Bušta J, et al. 2014. The Sketch Engine: ten years on. Lexicography, (1): 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff A, Jakubíˇcek M, Kováˇr V, et al. 2014. Finding terms in corpora for many languages with the Sketch Engine. EACL 2014.Lei Lei and Dilin Liu. 2016. A new medical academic word list: A corpus-based study with enhanced methodology [J]. Journal of English for Academic Purposes, (22): 42–53. https://doi.org/10.1016/j.jeap.2016.01.008

Longxing Li. 2023. The Keywords, Representation, and Conceptualization of China's Reform in the State Media Discourse: A Corpus-Assisted Critical Study. Doctoral dissertation, University of Macau.

Longxing Li, Chu-Ren Huang, and Xuefeng Gao. 2018. A SkE-Assisted comparison of three "prestige" near synonyms in Chinese. In J.-F. Hong, Q. Su, & J.-S. Wu (eds.), Chinese Lexical Semantics 19th Workshop. Springer, Cham. 256-266. https://doi.org/10.1007/978-3-030-04015-4_22

Longxing Li, Sicong Dong, and Xian Wang. 2020. *Gaige* and *reform*: A Chinese-English comparative keywords study. In Qi Su & Weidong Zhan (eds.), From Minimal Contrast to Meaning Construct: Corpus-based, Near Synonym Driven Approaches to Chinese Lexical Semantics. Springer, Singapore. 321-332. https://doi.org/10.1007/978-981-32-9240-6_22

Longxing Li, and Xian Wang. 2021. The development of COVID-19 Word List from the perspective of emergency language services. *China Terminology*, *23*(2), 32. https://doi.org/10.3969/j.issn.1673-8578.2021.02.005

Longxing Li, Xian Wang, and Chu-Ren Huang. 2021. Social Changes Manifested in the Diachronic Changes of Reform-related Chinese Near Synonyms. In Minghui Dong, Yanhui Gu, Jia-Fei Hong (eds.), Chinese Lexical Semantics. CLSW 2021. Cham: Springer. 184-193. https://doi.org/10.1007/978-3-031-06547-7_15

Saed, H., Hussein, R., Haider, A., Al-Salman, S., and Odeh, I. 2022. Establishing a COVID-19 lemmatized word list for journalists and ESP learners. *Indonesian Journal of Applied Linguistics, 11*(3), 577-588. https://doi.org/10.17509/ijal.v11i3.37103

West M. 1953. A General Service List of English Words. London: Longman.

WHO. 2024. Strategic Framework for Enhancing Prevention and Control of Mpox 2024–2027. Geneva: World Health Organization.