Public Health and Informatics
J. Mantas et al. (Eds.)
© 2021 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI210222

Evaluating Complexity of Digital Learning in a Multilingual Context: A Cross-Linguistic Study on WHO's Emergency Learning Platform

Yu ZHAO[a,1], Giuseppe SAMO[b], Heini UTUNEN[a], Oliver STUCKE[a] and Gaya GAMHEWAGE[a]
[a] *World Health Organization*
[b] *Beijing Language and Culture University*

Abstract. Reproduction of knowledge, especially tacit knowledge can be expensive during a pandemic. One of the most common causes is the reduced information accessibility during the translation process. Having the ability to assess the linguistic complexity of any given contents could potentially improve knowledge reproduction. Authors conduct two cross-linguistic studies on the World Health Organization (WHO)'s emergency learning platform to assess the linguistic complexity of two online courses in 10 languages. Morpho-syntactically annotated treebanks, unannotated materials from Wikipedia and language-specific corpora are set as control groups. Preliminary findings reveal a clear reduced complexity of learning contents in the most candidate languages while retaining the maximum amount of information. Creating a baseline study on low-resourced languages on the learning genre could be potentially useful for measuring impact of normative products at country and local level.

Keywords. Digital learning, health emergencies, COVID-19, linguistics, computational linguistics, Natural language processing

1. Introduction

Reproduction of knowledge, especially tacit knowledge [1], can be expensive during a pandemic. One of the most common causes is the reduced information accessibility during the translation process, in which an accurate adaptation of the technical complex contents such as medical information, requires huge amount of time and intensive labour. Having the ability to assess the linguistic complexity of any given learning contents in advance could potentially optimise the usage of translation resources, enabling a higher machine translation performance thus lowering the cost of a robust and economical knowledge reproduction.

[1] Corresponding Author, Yu Zhao, Technical officer, Department of Digital health and innovation, World Health Organization. Avenue Appia 20, Geneva, Switzerland; E-mail: zhaoy@who.int.

2. Materials and methods

To test the feasibility of the idea, authors conduct two cross-linguistic studies on the World Health Organization (WHO)'s emergency learning platform - Openwho.org. The dataset from the test group is extracted from two courses[2] in 10 languages, respectively Amharic, Chinese, English, French, Italian, Polish, Portuguese, Spanish, Russian, and Yoruba. Morpho-syntactically annotated treebanks (Universal Dependencies, [2]), unannotated materials collected in Wikipedia [3] and language-specific corpora are set as control groups. Both smart corpora and large-scale databases are used to provide qualitative and quantitative insights.

3. Findings[3]

In study 1, the WHO test corpus is compared with two datasets: (i) a multigenre, PUD (Parallel UD) for English, French, Portuguese, Russian, Spanish, Amharic ATT, for Amharic, Yoruba Treebank YTB4 for Yoruba and (ii) Wikipedia. The length of sentences is minimized in all investigated languages (binomial $p < 0.05$) with respect to OpenWHO, resulting in a smaller probability in having complementizers (and therefore complex structures). Only Amharic shows asymmetries with Wiki.

In study 2, syntactic complexity [4] is calculated on the basis of relativization strategies, resulting complex in cognitive studies and deep learning. We detect clear asymmetries between the distribution of relatives in the OpenWHO corpus compared to another parallel dataset (PUD, text genres: wiki, news) when transferring knowledge.

4. Conclusions

In future research, test dataset can include other under-resourced languages as a base for qualitative analyses to be merged within quantitative dimensions. Moreover, creating a baseline study on the learning genre could be useful for measuring impact of medical normative products, such as WHO's guidelines and guidance, at country and local level.

References

[1] David P, Foray D. An Introduction to the Economy of the Knowledge Society. International Social Science Journal. 2002;54:9-23.

[2] Nivre J. Towards a universal grammar for natural language processing. In International conference on intelligent text processing and computational linguistics; 2015 Apr 14; Springer, Cham: pp. 3-16.

[3] Eckart T, Quasthoff U. Statistical Corpus and Language Comparison on Comparable Corpora. In: Sharoff S, Rapp R, Zweigenbaum P, Fung P, editors. Building and Using Comparable Corpora. Berlin, Heidelberg: Springer; 2013. p. 151-65.

[4] Samo G, Zhao Y, Gamhewage G. Syntactic Complexity of Learning Content in Italian for COVID-19 Frontline Responders: A Study on WHO's Emergency Learning Platform, Verbum 11, 2020.

[2] Study 1 course: *Emerging respiratory viruses, including COVID-19: methods for detection, prevention, response and control*; study 2 course: *Infection Prevention and Control (IPC) for COVID-19 Virus*.

[3] Detailed datasets and visual displays of the findings can be retrieved via http://tiny.cc/wi3rtz.

[4] Info on text genres and sizes can be found https://universaldependencies.org/ (14 January, 2021).