Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210893

Could Linguistic Complexity Be Automatically Evaluated? A Multilingual Study on WHO's Emergency Learning Platform

Giuseppe SAMO[a,1], Yu ZHAO[b], Maria Teresa GUASTI[c], Heini UTUNEN[b], Oliver STUCKE[b] and Gaya GAMHEWAGE[b]
[a] *Beijing Language and Culture University*
[b] *World Health Organization*
[c] *University of Milano-Bicocca*

Abstract. The ability of assessing any type of linguistic complexity of any given contents could potentially improve knowledge reproduction, especially tacit knowledge which can be expensive during a pandemic. In this paper, we develop a simple and crosslinguistic model of complexity which considers formal accounts on the study of linguistic systems, but can be easily implemented by non-linguists' groups, e.g., communication experts and policymakers. To test our model, we conduct a study on a corpus extracted from the World Health Organization (WHO)'s emergency learning platform in 6 languages. Data extracted from open-access encyclopaedic entries act as control groups. The results show that the measurements adopted signal a trend for a minimization of complexity and can be exploited as features for (automatic) text classification.

Keywords. Digital learning; Linguistic complexity; Digital Health; Emergency training; Knowledge Reproduction

1. Introduction

Digital technologies revolutionized the way we retrieve information and acquire knowledge [1]. The recent COVID-19 pandemic has increased the demand for reliable information to help frontline health personnel respond to the outbreak, communities better protect themselves, and health policymakers draft new policies to accommodate emerging needs [2]. Being capable to apply distance learning became instrumental for health workers, who demonstrate different degrees of accessibility to content due to multilingual background, especially for non-native speakers. Under such circumstances, content providers are required to facilitate teaching and/or training activities in the context of emergencies, possibly in a variety of languages [3], which could be fairly time-consuming and labour intensive. Having the ability to assess the linguistic

[1] Corresponding Author, Giuseppe Samo, Department of Linguistics, Beijing Language and Culture University, Mailbox 82, Xue Yuan Road 15, 100083 Beijing, People's Republic of China; E-mail: samo@blcu.edu.cn.

complexity of any given contents in advance could potentially optimize resources and enabling the learning contents to be more readable/accessible [4].

Although indices of readability do exist, they are mainly language-specific (see [5] for an overview) and the language-specificity is not efficient to measure such indices across multiple languages. Moreover, many indices are also very coarse, generally on mean length of words and sentences.

Is there a way for us to quantify linguistic complexity from a crosslinguistic perspective? How can one observe and interpret trends of complexity, even without technical background in formal accounts of linguistics?

In this paper, we aim to discuss a simplified measurement of linguistic complexity that takes into account a simple intuition from the study of language in theoretical linguistics, psycholinguistics, computational linguistics, yet simultaneously appear to be easily implemented by non-linguists' group e.g., communication experts and policymakers.

The primary objective of this paper intends to operate on small-sized corpora, which are often under-investigated by the linguistic community, yet reveals critical value for under-resourced languages. Iterated from previous studies, we propose a model, an improved version from origin discussed and developed in [3] and [4], to minimize manual investigation for assessing complexity in health-related context.

We then continue discussing complexity measures in Section 2, and present an observational study in section 3, by comparing an emergency corpus from OpenWHO.org, World Health Organization (WHO)'s emergency learning platform with similar content from open-access encyclopaedic entries in six languages. In Section 4 we discuss and explore directions for further in-depth research, followed by a conclusion.

2. A simple model to measure linguistic complexity.

One of the objectives in theoretical and experimental linguistics is to investigate complexity in language comprehension and production, in adult and development grammars, and in language pathology (see [6] for an introduction and overview).

In experimental linguistics, different linguistic tasks are created to evaluate whether specific linguistic architectures (e.g., specific word orders, specific grammatical structures) are harder than others. Various types of measures such as reading (parsing) times, neuronal activity or, in specific populations, performances in specific tests (e.g., picture/scenario matching, see [7]) are defined for evaluation.

The evaluation results from experimental linguistics confirm a very basic idea originated from formal models of language: there are structures (e.g., a specific configuration of a sentence which opens to grammatical re-orderings of syntactic constituents) that are more complex to be parsed by human mind than others. Many of these syntactic re-orderings (e.g., relative clauses, topicalizations, cleft structures), which are also complicated for machine learning architectures in deep learning (see [8] for an overview), are cued by functional elements.

In Zipf's distributions [9, 10], functional elements have the tendency to be more frequent in corpora than lexical entries. The distribution of functional elements can be then detected by measures such as Gini's coefficient and Shannon *H* entropy (see 10 [for] a detailed discussion). We will not discuss elements Type/Token ratio (TTR), as content-specific text (e.g., medical information) require also the occurrences of technical terms. The role of measurements such as TTR, however, could be investigated in future studies.

Our main hypothesis is that in less complex texts, functional elements should be minimized, leading to an expected lower Gini's coefficient and higher Shannon entropy correlates if a text of the same size and similar TTR is compared.

3. A crosslinguistic study of learning contents from OpenWHO.org

We test our hypothesis on an emergency learning corpus, in which the materials have been well studied (see [3] and [4]) and labelled as less complex to other corpora (encyclopaedic entries, legal, news and social media) *via* a manual linguistic analysis on a typology of functional elements. All the materials (languages and size of the corpora are to be found in Table 1) are extracted from OpenWHO.org [11]. The dataset from the test group is extracted from the course *Infection Prevention and Control (IPC) for COVID-19 Virus* (downloaded in May 2020) in six languages: English, French, Italian, Polish, Portuguese and Spanish. For the control group, due to the transferring-knowledge nature of the dataset (see the discussion in [3]), we collected content from open-encyclopaedic entries (Wikipedia) on Covid-19 in the relevant six languages.[2] To avoid bias on the measures (see [10]), we extracted similar number of tokens, guaranteeing a readability of the content.

We compiled all the relevant textual contents in a *.txt* file and uploaded them to Zipfsexplorer (https://zipfexplorer.herokuapp.com/zipfexplorer, [10]). The online tool provides relevant measures discussed in Table 1.

The results in Table 1 support our hypothesis. All the Shannon's entropy measures are higher, while all the Gini's coefficients are equal or lower, as predicted by the hypothesis.

Table 1. Languages, Set (Wikipedia, OpenWHO.org), size, TTR (Type/Token ratio), Gini's coefficient and Shannon *H* entropy for every corpus under investigation (https://zipfexplorer.herokuapp.com/zipfexplorer. 08/2021). Results in bold confirm our hypothesis.

Language	Set	Size (Tokens)	TTR	Gini	*H*
English	Wiki	5,621	0.14	0.70	3.34
English	OpenWHO	5,087	0.15	**0.69**	**3.36**
French	Wiki	6,845	0.12	0.76	3.15
French	OpenWHO	6,886	0.11	**0.74**	**3.48**
Italian	Wiki	6,713	0.14	0.73	3.18
Italian	OpenWHO	6,646	0.12	**0.71**	**3.56**
Polish	Wiki	6,076	0.19	0.65	3.08
Polish	OpenWHO	5,994	0.18	0.65	**3.13**
Portuguese	Wiki	6,582	0.12	0.73	3.44
Portuguese	OpenWHO	6,591	0.12	0.73	**3.46**
Spanish	Wiki	7,960	0.10	0.78	3.36
Spanish	OpenWHO	8,042	0.10	0.78	**3.44**

[2] The control group data are extracted from the relevant URLs (last access, August 2021):
English: https://en.wikipedia.org/wiki/COVID-19;
French: https://fr.wikipedia.org/wiki/Maladie_%C3%A0_coronavirus_2019;
Italian: https://it.wikipedia.org/wiki/COVID-19; Polish: https://pl.wikipedia.org/wiki/COVID-19;
Portuguese: https://pt.wikipedia.org/wiki/COVID-19; Spanish https://es.wikipedia.org/wiki/COVID-19.

4. Conclusions

To conclude, we attempt to evaluate linguistic complexity of digital learning contents from OpenWHO.org, WHO's health emergency learning platform, across multiple languages.

We decide to apply an improved model to largely reduce manual investigation and provide possibility for non-linguistic experts to independently measure the complexity of health information prior to translation and information retrieval for training or monitoring purpose.

Finally, findings in this paper reveal potential possibility of further studies in different language families to test the reliability of such hypothesis. If adequately demonstrated, such model might be applied and developed as an additional feature for automatic text classification or text detection, contributing to a better machine learning performance in information retrieval and machine translation.

References

[1] Foray D. Economics of knowledge. Cambridge, MA: MIT press; 2004. 287 p.
[2] Gamhewage G, Utunen H, Attias M, George R. Fast-tracking WHO's COVID-19 technical guidance to training for the frontline. Weekly Epidemiological Record. 2020, Jun;95: 257-64.
[3] 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. 2020, Dec; 11:1-4.
[4] Zhao Y, Samo G, Utunen H, Stucke O, Gamhewage G. Evaluating Complexity of Digital Learning in a Multilingual Context: A Cross-Linguistic Study on WHO's Emergency Learning Platform. Stud Health Technol Inform. 2021, May; 281: 516-7.
[5] Dell'Orletta F, Montemagni S, Venturi G, READ-IT: assessing readability of Italian texts with a view to text simplification. In: Alm N, editor. SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, 2011 Jul 30; Edinburgh, UK. Stroudsbourg, PA, USA. p. 73-83.
[6] Guasti MT. Language acquisition: The growth of grammar. Cambridge, MA: MIT press; 2017. 672 p.
[7] Friedmann N, Belletti A, Rizzi L. Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. Lingua. 2009, Jan; 119: 67-88.
[8] Linzen T, Baroni M. Syntactic structure from deep learning. Annual Review of Linguistics. 2021, Jan; 7: 195-212.
[9] Zipf GK. Human Behavior and the Principle of Least Effort. Cambridge, MA: Addison-Wesley Press; 1949. pp. 573
[10] Coats S. Comparing word frequencies and lexical diversity with the ZipfExplorer tool. In; Reinsone S, Skadina I, Baklāne A, Daugavieti J, editors. Proceedings of the 5th Digital Humanities in the Nordic Countries Conference; 2020 Oct 21-23; Riga, Latvia. Aachen, Germany: CEUR, c2020. p. 219-25.
[11] Rohloff T, Utunen H, Renz J, Zhao Y, Gamhewage G, Meinel C. OpenWHO: Integrating Online Knowledge Transfer into Health Emergency Response. In: Dimitrova V, Praharaj S, Fominykh M, Drachsler H, editors. EC-TEL Practitioner Proceedings 2018: 13th European Conference on Technology Enhanced Learning; 2018 Sep 3-6; Leeds, UK, CEUR-WS, c2018. p. 1-14.