D4.9 Evaluation and assessment of methods for crowdsourcing in lexicography

Authors: Miloš Jakubíček

Date: July 31st, 2022

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

european lexicographic infrastructure

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D4.9 Evaluation and assessment of methods for crowdsourcing in lexicography

Deliverable Number: D4.9
Dissemination Level: Public
Delivery Date: July 31, 2022
Version: 1
Authors: Iztok Kosem, Miloš Jakubíček

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
1, July 31st	Submission-ready	Miloš Jakubíček

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

1 Introduction

This document contains a summary of the evaluation of the methods for crowdsourcing of lexicographic content which were developed and described in the Deliverable D4.3 Crowdsourcing Module in the ELEXIS infrastructure. The Crowdsourcing Module is described in Figure 1 below.

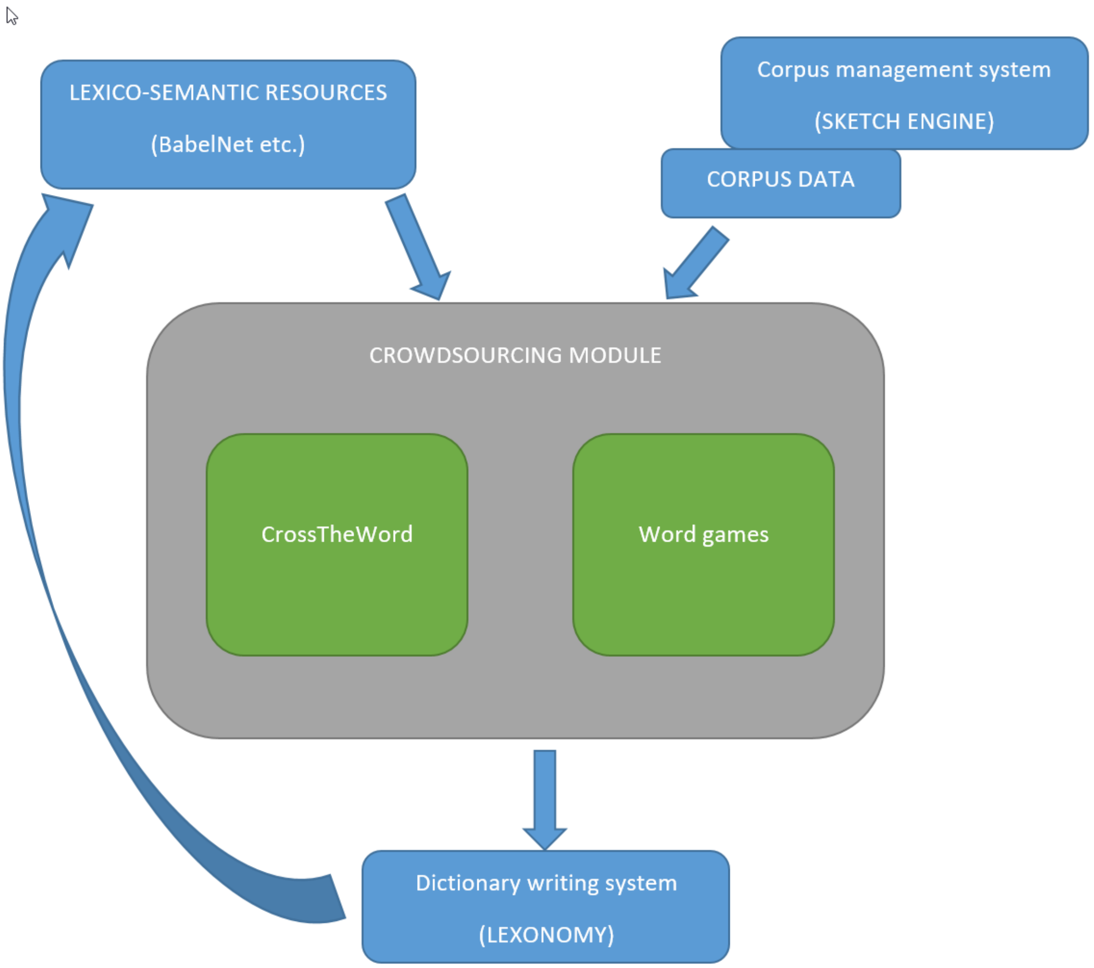

Figure 1: Crowdsourcing module in ELEXIS.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2 CrossTheWord mobile app

CrossTheWord is an innovative puzzle video game with a purpose for Android mobile phones, developed by Uniroma1 and based on components made available by the Babelscape Sapienza spin-off company. The code and documentation is publicly available on GitHub: https://github.com/elexis-eu/CrossTheWord.

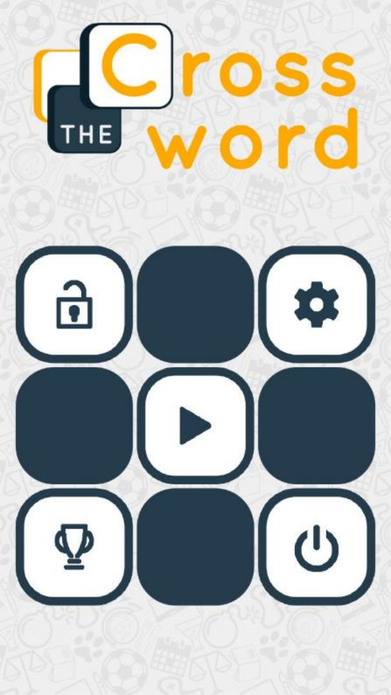

Figure 2: Crossword mobile application

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3 Word games mobile app

Word games is an innovative video game with a purpose focussed on word combinations. It is available for Android and iOS mobile devices. On Android devices, the users can register using a Google account, whereas on iOS the registration is made through Game center. The code and documentation is publicly available on GitHub: https://github.com/elexis-eu/word-games. On both platforms, users are asked to provide two additional types of information relevant for data analyses: age range and native language. At the moment, the game is available for Dutch, English, Estonian, Portuguese, and Slovenian, with other languages to be added later on. The users are expected to be both native and non-native speakers of languages offered.

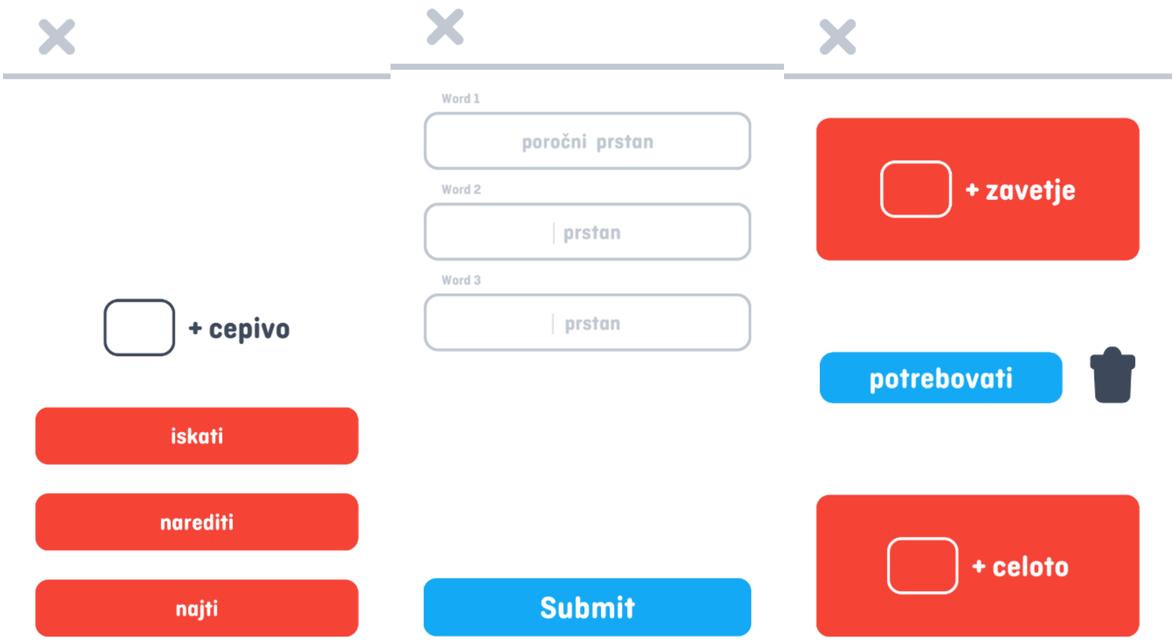

Figure 3: Three playing modes of the Word Game: CHOOSE, TYPE, DRAG (from left to right, Slovenian version).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4 Assessment and Conclusions

The Word games have been thoroughly described and assessed in publications [1], [2] and [3]. There is also significant previous work described e.g. in [4] and [5]. In general, there is no doubt that crowdsourcing can be a beneficial process in lexicography in terms that non-professional native speakers (and sometimes also non-native speakers) can contribute to create lexicographic data, either intentionally or unintentionally through some kind of games with a purpose or by validating existing or new (automatically generated) lexicographic content.

However, it still remains to be investigated how to motivate large crowds to take part in any such actions: it is a key idea in crowdsourcing that there are multiple (ideally dozens of) judgments for each particular case to be judged (such as a collocate or synonym to be assessed, a dictionary example to be validated or sense definition to be provided). In the context of lexicography, the necessary crowd needed are thus dozens of thousands of individuals.

While the theoreticaly aspects as well as practical implementations can be provided, scalability of the crowd that would enable the approach to be used in lexicographic production remains to be a standing issue.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

References

[1] ARHAR HOLDT, Š., et al. “Game of Words”: Play the Game, Clean the Database. *EURALEX XIX*, 2021.

[2] KUHN, T. Z., et al. Crowdsourcing pedagogical corpora for lexicographical purposes. In: *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*. 2021. p. 771-779.

[3] PORI, Eva, et al. The attitude of dictionary users towards automatically extracted collocation data: a user study. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 2020, 8.2: 168-201.

[4] Cibej, J., Fišer, D., & Kosem, I. (2015). The role of crowdsourcing in lexicography.

[5] Cibej, J., & Holdt, Š.A. Repel the Syntruders! A Crowdsourcing Cleanup of the Thesaurus of Modern Slovene.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.