elexis european lexicographic infrastructure

D4.3

CROWDSOURCING MODULE

Author(s): Iztok Kosem, Federico Martelli, Roberto Navigli, Miloš Jakubíček, Jelena Kallas

Date: 31. 1. 2020

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D4.3 Crowdsourcing module

Deliverable Number:	D4.3
Dissemination Level:	Public
Delivery Date:	31. 1. 2020
Version:	1.0
Author(s):	Iztok Kosem Federico Martelli Roberto Navigli Miloš Jakubíček Jelena Kallas

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
07/01/2020	First draft	Iztok Kosem
16/01/2020	Section on CrossTheWord added	Federico Martinelli
25/01/2020	Feedback	Roberto Navigli, Miloš Jakubiček, Jelena Kallas
31/01/2020	Final version	Iztok Kosem, Simon Krek

Table of Contents

List of Tables

List of Figures

1 Introduction

This document describes the Crowdsourcing Module in the ELEXIS infrastructure, which is currently formed of two games with a purpose that take lexico-semantic resources or corpus data as input for crowdsourcing.

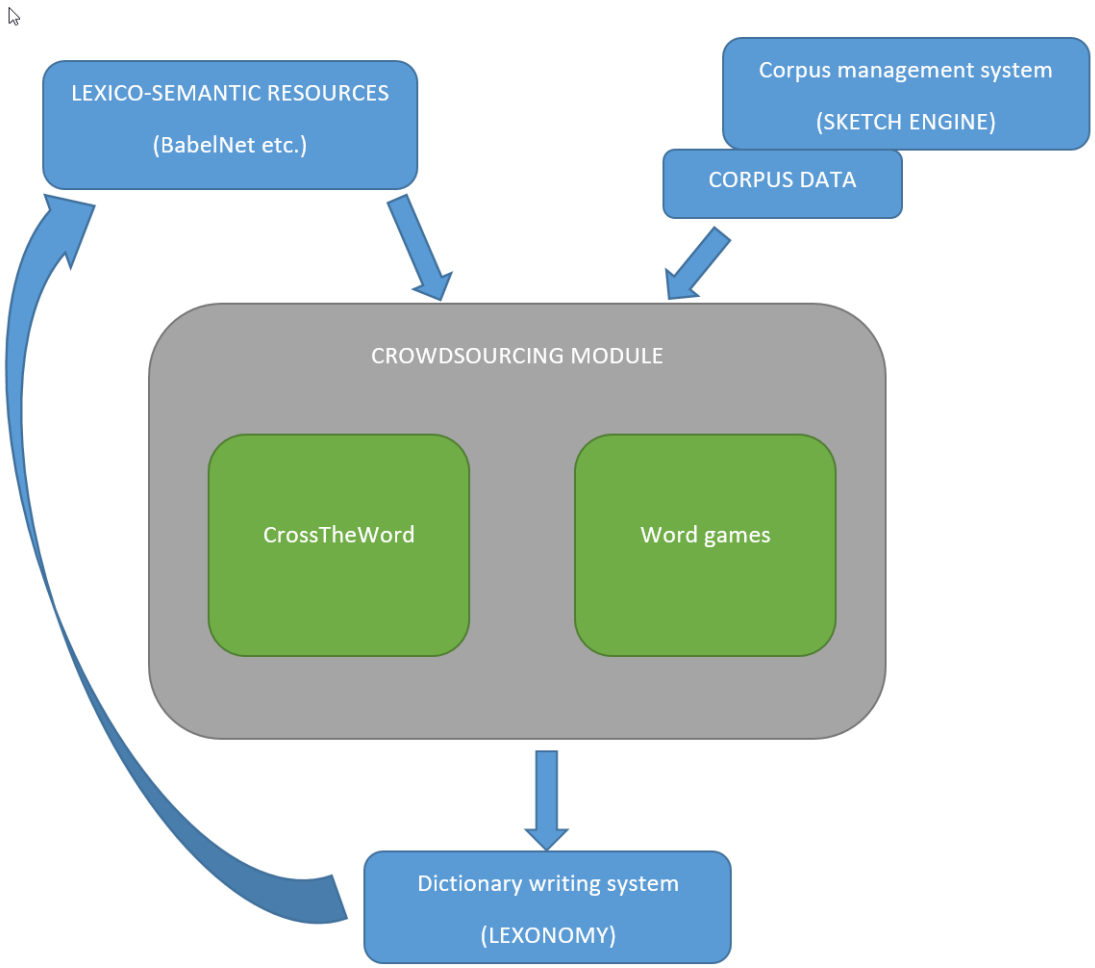

Figure 1: Crowdsourcing module in ELEXIS.

Addressing these two different avenues means that the games serve different needs of the developers of language resources: improving existing (automatically-generated) lexico-semantic resources, or cleaning corpus data before presenting it to the lexicographers. Interoperability with corpus management system and, due to the Dictionary drafting module, also with dictionary writing system

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.

are ensured (Figure 1), meaning that the information obtained from the games with a purpose can be easily integrated into the ELEXIS platform.

In the next sections, both games with a purpose are presented in more detail.

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.

The new user can register using the provided form or via their facebook account. After successful login, the user can either play a new crossword or continue a previous one.

Before starting a new crossword, the user is asked to select a degree of difficulty and the crossword language. Difficulty is related to both the crossword size and the lexical complexity of the crossword. As far as the language is concerned, in this first release, only English is provided, with other languages to be integrated later. Puzzle clues are generated starting from synset (i.e. meaning) definitions retrieved from WordNet, whereas answer words are senses of the corresponding synset. We report some examples of words, corresponding WordNet synsets and definitions in Table 1.

Figure 2: Main menu.

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.

Word	Synset	Definition
own	wn:02204692v	have ownership or possession of
being	wn:13954253n	the state or fact of existing
year	wn:15203791n	a period of time containing 365 days
one	wn:02186338s	used of a single unit or thing
son	wn:10624074n	a male human offspring
consonance	wn:04984351n	the property of sounding harmonious
charity	wn:04840405n	a kindly and lenient attitude toward people
nucleus	wn:09375085n	the positively charged dense center of an atom
wink	wn:00008435v	force to go away by blinking
corpus	wn:07955455n	a collection of writings

Table 1: Words, corresponding WordNet synsets and definitions extracted from WordNet.

The user can select a word to be guessed by swiping over the corresponding white cells of the puzzle as shown in Fig 1. Subsequently, the relating definition is shown and the user can use the keyboard to enter an answer. If the guess is correct, the user gains experience points, which can be used to unlock a bonus and solve more complicated clues.

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.

Figure 3: A crossword example.

Importantly, after guessing a word, a popup window is shown. Such window provides six randomly chosen potential substitutes (synonyms) extracted from BabelNet for the sense of the guessed word. The user is asked to select the correct substitutes according to the guessed word and its definition. In this way, it is possible to reduce noise in BabelNet which is an automatically created multilingual computational lexicon.

To determine the degree of synonymy between a given sense and a substitute selected by the user as correct, statistics are collected and stored in a database.

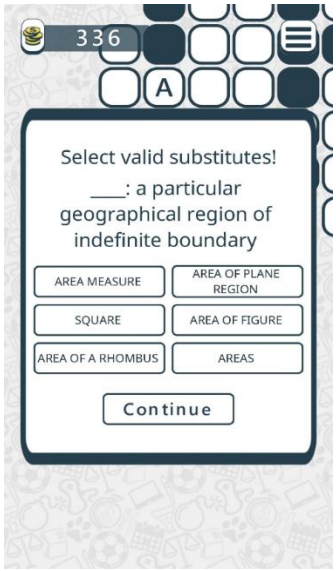

Figure 4: Pop-up validation window.

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.

2.1 Lexicographic value

The goal of CrossTheWord is twofold. First, it aims at improving the quality of lexico-semantic resources such as BabelNet, which play a crucial role in the ELEXIS lexicographic platform. Specifically, CrossTheWord leverages crowdsourcing for the purpose of removing noise within semi-automatically created resources. Second, once the ELEXIS dictionary matrix will be available, CrossTheWord will also be able to operate in arbitrary languages. This will enable interoperability between lexical resources both monolingually and across languages.

There are plans to include additional features to the game, such as additional popup windows (e.g. for the validation of similar words, collocations, etc.).

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.

The game draws on the collocational data, which is structured as follows:

- collocation id, grammatical relation name, headword position in the collocation, first word, second word, frequency, logDice value, weight

Currently, only binary grammatical relations are supported.

Users can choose from two game modes: Level-by-level and Practice. In the former, users start from Level 1, and proceed to the next level by completing all games in that level. Levels increase by difficulty, i.e. more difficult words and game modes are offered. In Practice mode, users are initially offered a randomly selected level, but are also able to choose any level themselves. The difference between the modes is that Practice does not count towards total score and user's position on the leaderboard.

Each level consists of up to 10 different games. Three types of games can be used:

- CHOOSE. In this type, the users are presented with three packages of three collocations and in each package they have to select the best/most typical one (typicality is measured according to logDice value). At the end, they are asked to order their three selected collocations according to their typicality, getting bonus points for the correct order. Collocations used in the game are selected from three ranges in the list of collocates, one from each range – top 30%, 30-55%, 55%-rest. In this way the cases where three collocates next to each other in the order are not selected.

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.

- TYPE. In this type, the users are provided with a headword and grammatical relation, and required to type up to three collocations. They get 0, 5, 20, 50 and 100 points, depending on the collocation position on the list.
- DRAG. In this type, the players need to drag/swipe provided words (collocates) to one of the three options: headword A, headword B, and Trash.

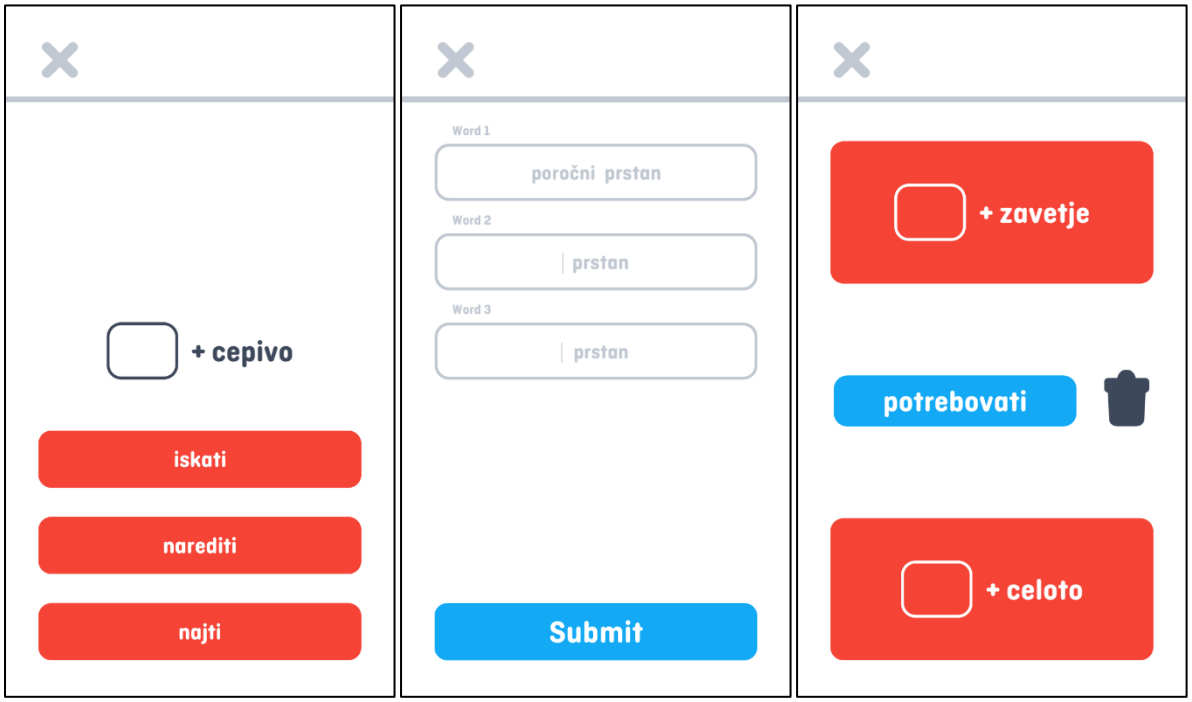

Figure 5: Three playing modes: CHOOSE, TYPE, DRAG (from left to right, Slovenian version).

A mixture of all three types is usually found in each level, although levels can contain only one game type. TYPE is deemed to be the most difficult of the game types and is thus used less frequently in initial (easier) levels, and more frequently later on.

Each game within level is scored, with bonuses such as double points for a game or a bonus word (giving bonus points) provided to users depending on their success, such as providing three or five correct answers in a row.

The Slovenian version has been developed further and includes a competition mode for Collocations, where users can play the same game against each other simultaneously, and Synonyms (solo mode). We look to provide similar modes, and new ones, for other languages in the future.

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.

3.1 Lexicographic value

Word games can be used to clean (semi)-automatically extracted collocational data from the corpus. With corpora getting increasingly larger, the abundance of collocational data is also increasing, putting heavy strain on lexicographers. It is known that semi-automatically extracted data contains different levels of noise, depending on the language and quality of tool such as lemmatizers, taggers and parsers. By playing the game, users can validate or reject collocation candidates. Every decision is logged and saved in the database. Moreover, the data is dynamic, i.e. collocated offered in the game are changed after being played (voted on) X number of times. The X number is variable and decided by each language data provider.

It is important to note that as far as collocations are concerned, gamification should be much more successful than direct (explicit) crowdsourcing approach given that deciding on whether something is or is not a collocation is difficult for non-linguists.

To ensure quick turnaround, data providers will be able to monitor user activity and download crowdsourced data via admin console, which will be made available soon. Moreover, uploading of new data to replace already (parts of) cleaned data will be possible. Traceability of data, e.g. by using external collocations ids, is ensured to make implementation of results into lexicographic workflow virtually seamless.

It is also important to ensure connectivity with other ELEXIS tools. The data input format for Game of words is derived from Sketch Engine automatic output of collocational data for a headword. This means that, using customised XSL transformation script, it will be later possible to conduct export of Sketch Engine data (or One-click dictionary data) from the Lexonomy dictionary writing system, and import it into the game, and later import the cleaned game data back into the dictionary writing system.

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.

There are plans to add other types of games later on, for example synonyms (already proven viable for Slovenian), and sense distribution of examples of collocations (successfully piloted by Kosem et al. 2017)[1].

[1] Kosem, I., Gantar, P., Krek, S. (2017). Sense menus in collocations dictionary of Slovene. *Electronic lexicography in the 21st century (eLex 2017): lexicography from scratch, book of abstracts.* Leiden: Dutch Language Institut; Brno: Lexical Computing; Ljubljana: Trojina Institute for Applied Slovene Studies, p. 43.

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.