INIKOL - Collocational Database for Learning Croatian as a Foreign Language

Goranka Blagus Bartolec[1], Gorana Duplančić Rogošić[2], Antonia Ordulj[3]

[1] Institute for the Croatian Language Zagreb
[2] Faculty of Economics, Business, and Tourism Split
[3] Faculty of Croatian Studies Zagreb

gblagus@ihjj.hr, gduplanc@efst.hr, antoniasvetic@gmail.com

Abstract

This paper describes the ongoing work on the INIKOL project - the development of a collocation database for learning Croatian as a foreign language. The main goal of the project is to contribute to easier mastery of collocations as fixed phrases in Croatian as a foreign language.

1 Introduction

Collocations, which are multi-word expressions (MWEs) with a fixed structure and meaning, are a challenge for non-native speakers of Croatian who have difficulty understanding and using them in terms of: 1. recognizing the elements of a collocation and understanding their meanings (see Ordulj and Naumoska-Giel, 2022; Goh, 2000; Graham, 2006), 2. recognizing the part of speech when selecting collocates and choosing the correct morphological form of the inflected word (Ordulj, 2018a; 2018b), 3. using the correct preposition with the appropriate case form of the noun, 4. linking individual words to other words within a phrase or sentence.

The number of students learning Croatian has been steadily increasing over the last two decades, so the need for high-quality manuals and an applied description of the Croatian language is also growing. The groups of students learning Croatian as a foreign language are extremely heterogeneous in terms of mother tongue, age, gender and previous knowledge; their motives for learning Croatian are also different. Foreigners who learn Croatian as non-native speakers come from different language areas: many come from Slavic countries as students of Croatian studies and the Croatian language, so the structure and vocabulary of Croatian are close to them, but they also come from other language areas where the structure of Croatian is unfamiliar and more difficult to learn. Many participants come from South America and are descendants of Croats who emigrated from Croatia at the beginning or middle of the 20th century. There are also many learners from other European countries who are in Croatia for professional or private reasons and want to learn Croatian. In recent years, the number of immigrants, i.e. workers from Asian countries who need to learn Croatian at a basic level, has also increased.

Textbooks and exercise books for learning and teaching Croatian as a foreign language as well as the more recent *Basic Croatian Grammar for Croatian Language Learners* (Matovac, 2022) created in Croaticum, the largest institution for teaching, research and description of Croatian as a second and foreign language, offer non-native speakers a good insight into the structure and lexical potential of the Croatian language from beginner to intermediate level.

Given the frequency of collocations in everyday use and all the above-mentioned challenges that non-native speakers face when learning Croatian, the INIKOL project was developed to build a collocation database as an additional, publicly accessible resource that non-native speakers can use as an online tool for

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 8–12

searching, understanding and applying these expressions when learning Croatian and when communicating in Croatian in various contexts of use.

2 About the INIKOL database

The INIKOL database is part of two larger projects: MWE-Cro: Multiword Expressions in Croatian - Lexicological, Computational Linguistic and Glottodidactic Approach, and VIBA: Database of Croatian MWEs, which aim to build a complete, publicly accessible online platform for multi-word expressions in Croatian, which will include: a) several monolingual databases, namely of idioms, proverbs and multi-word expressions in general use and in languages for specific purposes, b) a multilingual database of verb collocations and c) a multilingual database of collocations in Croatian as a foreign language (INIKOL). Acquiring and understanding fixed word combinations in Croatian, as in other languages, is a challenging and demanding task for non-native speakers, especially for non-Slavic ones. The morphology of inflected nouns and verbs and the selection of the appropriate collocation are often a significant obstacle that makes it difficult for non-native speakers to learn and reproduce MWEs in Croatian. Therefore, the basic aim of building INIKOL is to contribute to the easier acquisition and use of collocations in Croatian as a foreign language. During the four-year project period[1], a total of 800 to 1,000 collocations from the basic vocabulary of Croatian as a foreign language will be entered into the INIKOL database. The collocations entered into the database follow the *Croatian A2: Descriptive Framework of Reference Level A2* (Grgić and Gulešić Machata, eds., 2017), which is based on the Common European Framework of Reference for Languages (CEFR) and the content of textbooks for teaching Croatian as a foreign language at level A2. All collocations are grouped into thematic areas (e.g. *MAN, EDUCATION, LIVING, TRAFFIC AND TRAVEL, FOOD AND DRINK, SHOPPING, SERVICES*). The main entries in the database are nouns (e.g. *family, school, house, train, city, language, sea, park, glass*...) and verbs (e.g. *to be, to go, to eat, to drink, to write, to study*, ...). Verbal and noun entries are key words under which collocations are listed in INIKOL. For example, the collocation *biti gladan*, eng. 'to be hungry,' is listed under the verb *biti*, eng. 'to be' as a verb entry, and the collocation *obiteljska kuća*, eng. 'family house' is listed under the noun *house* as a noun entry. Other parts of speech, such as prepositions, adverbs, and adjectives, will also be included as entries in the INIKOL in the following phase.

2.1 INIKOL structure: user interface outline

Collocations are entered in the online working interface (backend) and all entered data is visible in the user interface (frontend) after saving. The interfaces were created by an external IT developer according to the ideas of the project member (see Figure 1). The project members access the user interface (backend) via a user name and password, and while working on the project, the user interface (frontend) is only visible and accessible to the project members, but not publicly visible and searchable.

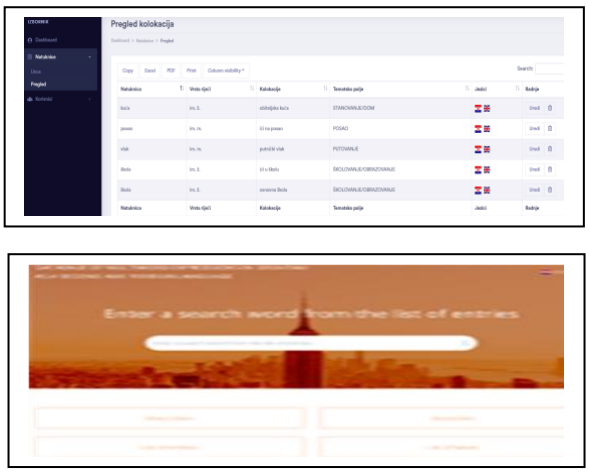

Figure 1: INIKOL – backend and frontend

The frontend of the INIKOL database, i.e. the final result of the project, will be publicly accessible and searchable under the Croatian domain jezik.hr after the project is completed.

When the interface opens to the public, users will be able to search: a) by individual entries (one-word lexemes), which are elements of multi-word expressions in Croatian and English; b) by multi-word expressions in Croatian (by selecting from a

[1] INIKOL is an ongoing project that will run from December 2023 to December 2027 at the Institute for the Croatian Language in Zagreb.

drop-down menu, which will be sorted alphabetically by the initial word of the multi-word expression); and c) by the thematic area to which the multi-word expression belongs. At the moment, it is only possible to search for Croatian and English entries in the test version of the frontend.

The user interface for INIKOL consists of seven fields (see layout of the backend interface in Figure 2). For Croatian, four fields are filled in: 'Entry' (one-word lexeme as collocation element under which collocations are entered in INIKOL), 'Part of speech' (for the entry); 'Collocation/MWE', 'Thematic field' to which the collocation belongs, 'Examples from the corpus' and three fields for English equivalents: 'Entry', 'Collocation/MWE' and 'Thematic field'.

Figure 2: Layout of backend interface in Croatian and English for collocation *ići u školu*, eng. 'go to school')

2.2 Corpus-based examples in INIKOL

An essential part of the development of the INIKOL database is the inclusion of example sentences from the corpus. Two or more example sentences are provided for each entry in order to confirm the use of the collocations in practice. The source corpus is the Croatian corpus *MaCoCu Croatian Web v2 2021–2022* (Bañón et al., 2023), which is available in Sketch Engine as the Croatian version of the multilingual corpus platform MaCoCu corpora, which were built by indexing the Internet's top-level domains in 2021 and 2022. In the Croatian version of MaCoCu, users can search for examples using the simple query option, with the possibility of searching for lemmatic forms of the elements of collocations. In this way, it is possible to find: 1. all paradigmatic forms in which collocations are recorded in the corpus, which is important because Croatian is an inflectional language, 2. the valency patterns between words at the syntagmatic and syntactic levels. After obtaining the overall results through a simple query, the GDEX option is selected for the automatic selection of sentences that are easily understandable for non-native speakers and suitable for teaching. Figure 3 shows the results of a corpus search for the collocation *ići u školu*, eng. 'go to work' in MaCoCu using the GDEX option.

Figure 3: The first 12 examples of the MWE *ići na posao*, eng. 'to go to work', retrieved using GDEX

Based on the retrieved examples, the entry is entered and edited in the backend by selecting sentences from the corpus that a non-native speaker could understand and that include different morphological forms of the individual elements of the collocation. The use of Croatian prepositions in certain cases is rather challenging for non-native speakers as the noun used determines the choice of preposition, e.g. the prepositions *u* and *na* (literally *in* and *on* respectively in English) are in Croatian '*ići na posao*', eng. 'to go to work' and *ići u školu*, eng. 'to go to school'. The inclusion of such collocations in INIKOL will make it easier for non-native speakers to use such collocations correctly in Croatian.

2.3 INIKOL database frontend

The frontend of the INIKOL database follows the structure of the backend interface and enables the display of entries in Croatian and English interfaces. Figure 4 shows the search for the entry *school* in the English interface. The user will be able to use both the Croatian and the English interface. The English equivalents will be retrievable through the English interface and the Croatian equivalents through the Croatian search engine.

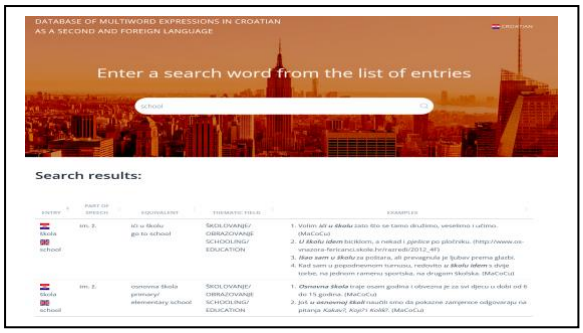

Figure 4: Layout of the INIKOL frontend with the entry *škola,* eng. 'school'

A list of the entries entered in INIKOL is available in Croatian and English on the home page of the frontend (see Figure 5) and can be searched using the Croatian or English search engine.

Figure 5: INIKOL English home page with a list of searchable and added entries

3 Further work on INIKOL

The INIKOL database is planned as a dynamic project that offers the possibility of updating existing data and entering new collocations according to user needs and changes in extra-linguistic reality. The database will be expanded to include additional fields and a sound recording of the pronunciation of each collocation.

Collocations at advanced levels will be added to the INIKOL database, which in the first phase includes the basic vocabulary, as the ultimate goal is to include collocations from all CEFR levels, with each collocation labelled with the level at which it is acquired. In addition to the English equivalents, the database will be expanded to include other languages from the countries from which non-native speakers who learn Croatian come. In a further step, in addition to the MaCoCu corpus, examples from the CroLTeC - CROatian Learner TExt Corpus (Mikelić Preradović et al., 2015) will be entered. CroLTeC contains essays collected from learners of Croatian as a second and foreign language (from A1 to C1 level).

The INIKOL project will provide online tools for Croatian as a foreign language and contribute to its visibility, prominence and use by non-native speakers. It will also help place Croatian in line with other European languages that already have such tools in wider use such as English (e.g. Oxford Collocations Dictionary; Collins Dictionary), Spanish (Dictionary of Collocations - DICE) or French (Le Robert).

Acknowledgments

This work was supported by the Croatian Science Foundation under the project number HRZZ-IP-2022-10-7697, and under the project VIBA - Database of Croatian MWEs funded by the EU - NextGenerationEU. Author(s) opinions do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.

References

Ana Grgić and Milvia Gulešić Machata (Eds.). 2017. *Hrvatski A2: Opisni okvir referentne razine A2 [Croatian A2: Descriptive framework of reference level A2]*. FF press, Zagreb.

Antonia Ordulj. 2018a. *Kolokacije u hrvatskom kao inom jeziku: Uvid u receptivno i produktivno znanje imenskih kolokacija*. HSN, Zagreb.

Antonia Ordulj. 2018b. Analiza odgovora imenskih kolokacija kod neizvornih govornika hrvatskoga jezika, *Journal for Foreign Languages*, 10 (1): 133-153. https://doi.org/10.4312/vestnik.10.133-153.

Antonia Ordulj and Karina Naumoska-Giel. 2022. "Razumijete li me?" ili o slušanju u nastavi hrvatskoga kao inoga jezika. In *Od ucha do ucha.* 243-255. Uniwersytet im. Adama Mickiewicza w Poznaniu, Poznań.

Collins. 2024. *Collins Online Dictionary*. https://www.collinsdictionary.com.

Christine M. Goh 2000. A cognitive perspective on language learners' listening comprehension problems. *System*, 28(1): 55–75. https://doi.org/10.1016/S0346-251X(99)00060-3.

Darko Matovac. 2022. *Basic Croatian Grammar: For Croatian Language Learners*. HSN, Zagreb.

DiCE: Diccionario de Colocaciones del Español. Facultade de Filoloxía (Universidade da Coruña). https://www.dicesp.com/paginas.

Dictionnaire français gratuit – Dico en ligne Le Robert. Le Robert. https://dictionnaire.lerobert.com.

Marta Bañón et al. 2023. *Croatian web corpus MaCoCu-hr 2.0*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042,1.

Nives Mikelić Preradović, Monika Berać, Damir Boras. 2015. Learner Corpus of Croatian as a Second and Foreign Language. *Multidisciplinary Approaches to Multilingualism*. In K.Cergol Kovačević, S.L.Udier. Peter Lang, F am M.107-126.

Oxford collocations dictionary at Oxford learner's dictionaries. 2024. Oxford University Press.

Suzanne Graham. 2006. Listening comprehension: The learners' perspective. *System*, 34(2): 165–182. https://doi.org/10.1016/j.system.2005.11.001.

Textbooks and workbooks – Croaticum. https://croaticum.ffzg.unizg.hr/?page_id=1632.