Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Integrating Croatian into Concepticon: a Corpus-Based Frequency Mapping of Croatian Vocabulary

Anja Krišto
Faculty of Linguistic Sciences and Foreign Literatures
Università Cattolica del Sacro Cuore

This study presents a Croatian frequency-derived wordlist mapped to Concepticon concept sets, based on the most frequent nouns, verbs, and adjectives extracted from the hrWaC web corpus. The resulting dataset connects corpus-based Croatian vocabulary to Concepticon's cross-linguistic framework and includes lexicalizations from nine additional languages for each mapped item.

1 Introduction

Standardized lexical datasets have become essential for systematic comparison across languages. Lexibank and Lexibank 2 (List et al. 2022; Blum et al. 2025) are just one of the resources that have made it possible to combine wordlists from diverse sources and connect them through shared identifiers. Concepticon (List et al. 2016, List et al. 2025) is another example that provides stable, language-independent concept set identifiers which allow diverse concept lists to be linked across languages and research traditions. Both Concepticon and the related NoRaRe database (Tjuka et al. 2023) have grown steadily, but language representation remains uneven. Many languages, including Croatian, are currently absent as mapped languages.

Croatian is a South Slavic language with a strong linguistic tradition and several high-quality digital resources covering usage, frequency, and semantic structure within the language. For example, HR-CLARIN provides language resources and technologies intended to support research in the humanities and social sciences. On the other hand, Croatian Psycholinguistic Database (Peti-Stantić et al. 2021), provides detailed

Computer-Assisted Language Comparison in Practice
Volume 9, Number 1
URL: https://calc.hypotheses.org/9392 DOI: 10.15475/calcip.2026.1.6

Published on 08/06/2026

information about word properties including subjective frequency, age of acquisition, and concreteness for a large set of Croatian lemmas. These resources are useful for research within Croatian, but they are not linked to cross-linguistic infrastructures and do not use shared identifiers. This means that Croatian lexical data cannot easily be compared with data from other languages, or integrated into Concepticon or similar comparative frameworks.

This contribution maps a corpus-derived Croatian frequency list to existing Concepticon concept sets. Instead of starting from a curated concept list, as is typical for Concepticon contributions, the starting point here is the most frequent Croatian word forms, asking which of them correspond to existing concept sets. The result is a dataset based on actual language use that is at the same time linked to Concepticon's cross-linguistic framework.

2 Background

2.1 Concepticon and NoRaRe

Concepticon is particularly well suited for this kind of work because it already provides a large, structured inventory of cross-linguistic concept sets that Croatian vocabulary can be mapped onto directly. Rather than building a new comparison framework, mapping Croatian to Concepticon connects it to an infrastructure that already links over 400 concept lists from 56 languages. Its limitations are also worth noting here. Concepticon is not a full formal ontology, as its semantic relations are designed for navigation and exploration rather than to encode strict semantic truths (Tjuka et al. 2023). For the present work, this means that mapping decisions do not need to resolve every ambiguity. The goal is a plausible, defensible link between a Croatian form and an existing concept set, not to make universal claims about meaning. Coverage is a more significant limitation: because Concepticon was built primarily from typological and historical lists, contemporary and domain-specific vocabulary is often absent from its concept inventory. This structural gap defines what the mapping can and cannot show, and is why the unmapped portion of the dataset is treated as a finding rather than a failure.

NoRaRe (Tjuka et al. 2023) extends Concepticon with psycholinguistic and cognitive data, in the form of norms, ratings, and semantic relations drawn from published experimental studies. The two resources are linked through the Concepticon ID, meaning that any concept present in Concepticon can be associated with additional information from NoRaRe. Adding Croatian to Concepticon therefore also opens a pathway toward future integration with NoRaRe.

2.2 Frequency-Based Approaches and Learner-Oriented Motivation

Frequency lists derived from large corpora are widely used in vocabulary research and language teaching because they provide empirical evidence of which words speakers most commonly encounter in real texts (Kilgarriff 2010). A small number of high-frequency items typically covers a large proportion of running words in texts (Nation 2001), making frequency a reasonable starting point for vocabulary selection.

For learners, though, a frequency list alone is not enough. It shows which words are common, but not how they relate to each other in meaning. Working with frequency-based approaches also reveals inherent limitations: while frequency lists record how often a form appears, they do not specify what it means in a given context. A single frequent form may have several distinct meanings, and without context there is no way to know which one is intended (Fulgosi and Tuđman Vuković 2001). Linking high-frequency items to concept sets helps with this by grouping synonyms and near-synonyms under a shared concept, making polysemy visible, and allowing multilingual comparison that goes beyond just translation pairs (Krišto 2026). Recent research has also argued that wordlist development has been dominated by English and a small set of well-resourced languages, and has called for broader, language-diverse approaches (Dang and Webb 2025). The Croatian frequency list presented here is a step in this direction.

Within Croatian linguistics specifically, frequency-based and corpus-based resources exist but remain language-internal in their design. The Croatian web dictionary Mrežnik (Hudeček and Mihaljević 2019), for example, shows how corpus evidence can support curated lexicographic resources, but it is not designed for cross-linguistic concept alignment or interoperability with infrastructures such as Concepticon.

3 Materials and Methods

3.1 Materials

The dataset was compiled from hrWaC (Croatian Web Corpus), a large collection of texts from Croatian web domains reflecting contemporary written usage on the internet (Ljubešić and Erjavec 2011), chosen because it is large, freely accessible, and covers a broad range of contemporary language. The 1,000 most frequent nouns, verbs, and adjectives were then extracted using Sketch Engine, giving an initial list of 3,000 items. These three word classes were chosen because Concepticon contains concept sets across all three categories, and working with only one would have left a large part of the available concept inventory unexplored.

3.2 Data Cleaning

After extraction, the raw wordlists were manually reviewed and cleaned. Several types of items were removed: proper names (such as people, cities, and nationalities), acronyms and administrative terms with no stable Concepticon equivalent, highly specific religious terms (such as names of holidays and religious figures), and items misclassified by the corpus tagger, such as verbs appearing on the adjective list. General terms such as *crkva* "church" or *bog* "god" were kept, while strongly culture- or institution-specific items were excluded. Native speaker intuition played an important role in this step, as it helped determine whether items were valid lexical entries suitable for later mapping. The cleaning reduced the dataset from 3,000 to 2,668 items: 981 verbs, 923 nouns, and 764 adjectives.

3.3 Translation into English

Translating the Croatian wordlists into English was a necessary step because Concepticon uses English as its meta-language for concept glosses, labels, and associated metadata. The translation combined online tools (Google Translate) with physical and digital dictionaries for cases where automated tools gave unclear results. All translations were manually checked by the author, a native Croatian speaker with advanced English proficiency, after which a second reviewer, a linguist with a degree in English, verified the translations for consistency and accuracy.

Translation and mapping turned out not to be fully separable stages. In several cases, consulting Concepticon's concept definitions led to changes in earlier translation choices, because the definitions made clear that a particular English gloss had a narrower or broader meaning than initially assumed. This kind of back-and-forth is difficult to avoid when working with a polysemous source language and a concept inventory built around specific meanings.

3.4 Automated Mapping with PyConcepticon

The mapping process combined automated methods with manual verification. As a first step, Python and the `PyConcepticon` library (Forkel et al. 2024) were used to generate initial candidate mappings. The tool takes English glosses as input and returns a single best-match Concepticon concept set per gloss. Because Concepticon concept sets are indexed via English labels, the quality of automated mapping depends directly on the quality of the translations produced in the previous step. The automated suggestions provided a useful starting point but required systematic correction. The tool does not perform sense disambiguation and therefore cannot resolve context-dependent meanings. Synonymy and polysemy also required attention: several Croatian words

sometimes pointed to the same concept, and a single Croatian word could have more than one English translation, not all of which had a match in Concepticon. For these reasons, all automated mappings were manually reviewed.

3.5 Manual Mapping and Mapping Categories

Words that were not automatically matched, or whose automated suggestions seemed uncertain, were mapped manually. This relied on a combination of native speaker intuition, bilingual dictionaries, and, where possible, comparisons with Concepticon mappings in related languages. During the annotation process, items were organized in four working categories: automatically mapped and verified (`PyConcepticon`), manually mapped, uncertain (many of which were later resolved), and unmapped. These categories are not retained in the final dataset but were useful for managing the workflow, especially for difficult and borderline cases. All mappings were carried out using Concepticon Version 3.4.0.

3.6 Adding Frequency Information and Multilingual Equivalents

After the mapping stage, corpus frequency information was added to each item in the dataset using Sketch Engine's built-in frequency tool. Each entry includes its absolute corpus frequency, its rank in the cleaned list, and the rank of its associated Concepticon concept set where applicable. This means that each mapping can be read alongside how frequently the word actually appears in contemporary Croatian.

As a final step, lexicalizations from nine additional languages were added for each mapped item by consulting the corresponding Concepticon entries: Serbian, Hungarian, Czech, German, Italian, French, Spanish, Russian, and Ukrainian. These languages were selected because they are regionally relevant, because some are spoken as minority languages, and because they are the languages most commonly taught in Croatian schools (Kapović 2022). This gives the dataset a multilingual dimension that goes beyond Croatian-English glossing and makes it more useful for cross-linguistic comparison.

4 Analysis

4.1 Mapping Coverage

Of the 2,668 items in the cleaned dataset, 1,262 (47.3%) were mapped to at least one Concepticon concept set, covering 856 unique concept set identifiers (Figure 1). In addition, 12 items were assigned two Concepticon IDs, so the dataset contains 1,274 total Concepticon assignments (mapping links). The remaining 1,406 items (52.7%) remained unmapped.

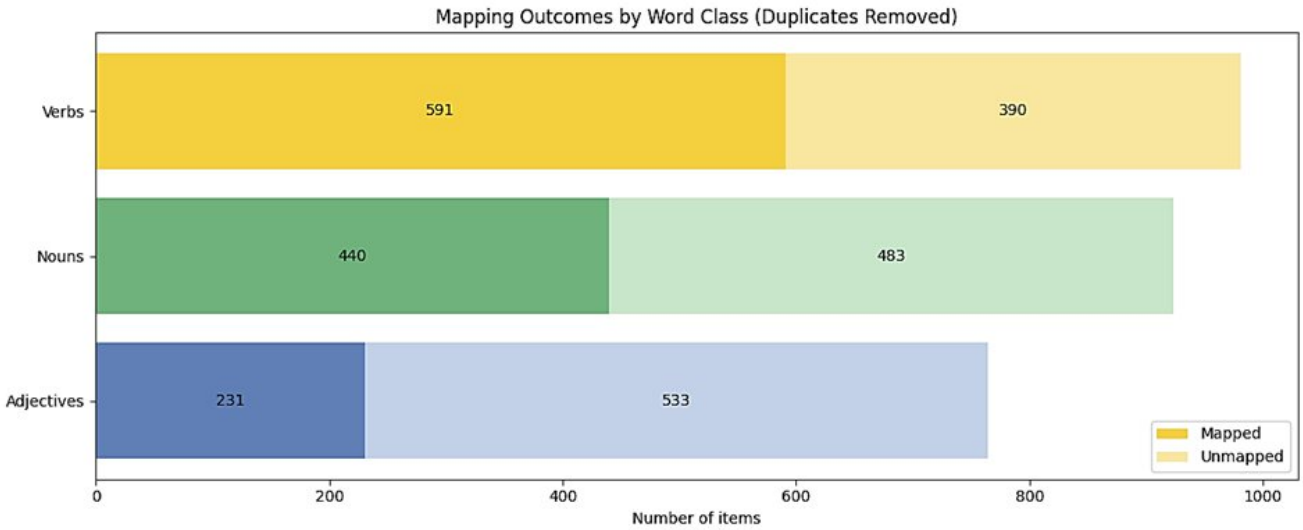

Figure 1: Mapped and unmapped items by word class. Verbs show the highest mapping rate, adjectives the lowest.

4.2 Unmapped Items: Derivational Mismatches and Modern Vocabulary

The majority of unmapped items fall into two broad patterns. The first one involves derivational and part-of-speech mismatches. Many unmapped Croatian items are semantically close to an existing concept set but appear in a derived form (as relational adjectives, agent nouns, or nominalized verbs) that does not match the lexical category Concepticon most commonly represents. Relational adjectives are particularly frequent in this group: forms such as *ljubavni* "love-related", *školski* "school-related", and *vjerski* "religious" are close in meaning to concept sets such as LOVE, SCHOOL, and RELIGION, but mapping them would mean collapsing the "related to X" meaning into the base noun concept, which was avoided in order to keep the mappings consistent and transparent.

The second pattern concerns modern institutional vocabulary. A large number of unmapped items belongs to domains that are prominent in web-based language use but largely absent from Concepticon's concept inventory: technology (e.g., web, online, portal), sports (e.g., football, fan, champion), politics and governance (e.g., politician, parliament, coalition), finance and money (e.g., paycheck, investment, budget), religion (e.g., believer, bishop, parish), and media and culture (e.g., art, history, cinematic). These clusters were identified through manual coding, grouping items by shared topic. This reflects a structural difference between the two resources: Concepticon was built primarily from lists motivated by typological and historical goals, where the focus is on stable, cross-culturally recurrent meanings, while a web corpus naturally surfaces vocabulary from contemporary public and institutional life.

A smaller group of unmapped items points to a different kind of problem: not a mismatch in word class or domain, but a mismatch between a universal concept and a

culturally specific meaning. For example, *dvojica* „two male persons“ could not be mapped to TWO because Concepticon's concept set does not encode the human and gender-specific restriction that is central to the Croatian form. These cases are small in number but illustrative: they show that some Croatian words carry cultural or grammatical specificity that sits outside what a language-independent concept inventory can currently represent.

4.3 Mapped Items: Many-to-One Clusters and Polysemy

Within the mapped subset, the most notable pattern is many-to-one mapping, where multiple Croatian forms linked to a single Concepticon concept set. This happens most often with near-synonyms, aspectual verb pairs, and prefix-derived verb variants. The concept set UNDERSTAND (ID 1536) attracted the largest cluster, with six Croatian forms: *kužiti*, *razumjeti*, *shvatiti*, *shvaćati*, *skužiti, and uvidjeti*. These forms share a conceptual core but differ in register, aspect, and usage context, which remain visible at the form level even though Concepticon groups them under one concept. Similar patterns appear with descriptive adjectives such as *velik*, *ogroman*, *golem*, and *krupan* all mapping to BIG (ID 1202), and with verb families such as *završiti*, *završavati*, and *dovršiti* all mapping to FINISH (ID 1766). Table 1 gives an overview of where this clustering is most pronounced.

Rank	Concepticon ID	Concept name	# Croatian forms	Croatian forms
1	1536	UNDERSTAND	6	kužiti, razumjeti, shvatiti, shvaćati, skužiti, uvidjeti
2	3004	FORMER	6	bivši, dosadašnji, nekadašnji, prethodan, prijašnji, tadašnji
3	7	GATHER	5	okupiti, okupljati, prikupiti, prikupljati, skupiti
4	680	BREAK (DESTROY OR GET DESTROYED)	5	probiti, puknuti, pući, razbiti, slomiti
5	692	BRING	5	donijeti, donositi, dovesti, dovoditi, ponijeti
6	707	REMEMBER	5	pamtiti, prisjetiti, sjetiti, sjećati, zapamtiti
7	1309	ASK (INQUIRE)	5	pitati, upitati, zamoliti, zapitati, zatražiti
8	1529	LAST (FINAL)	5	finalan, konačan, krajnji, posljednji, zadnji
9	1752	LEAVE	5	odlaziti, ostaviti, ostavljati, otići, prepustiti
10	1778	APPROACH	5	bližiti, približavati, pristupati, pristupiti, prići

Table 1. Top 10 Concepticon concept sets by number of mapped Croatian forms, with Croatian variants listed.

A small number of items, 12 in total, were assigned two Concepticon IDs rather than one, for different reasons. The clearest group consists of genuinely polysemous Croatian forms, where one word covers two conceptually distinct meanings. Several of these reflect well-known cross-linguistic patterns: *mjesec* maps to both MONTH (ID 1370) and MOON (ID 1313), *jezik* to both LANGUAGE (ID 1307) and TONGUE (ID 1205), and *zemlja* to both COUNTRY (ID 1300) and EARTH/SOIL (ID 1228). A second group reflects Concepticon's finer-grained sense inventory rather than polysemy in Croatian itself. *Dan*, for example, was mapped to both DAY (NOT NIGHT) (ID 1225) and DAY (24 HOURS) (ID 1260) because Concepticon distinguishes two senses that a context-free lemma cannot. Similarly, *kraj* and *završetak* were linked to both END (OF SPACE) (ID 742) and END (OF TIME) (ID 743). In all these cases, allowing two mappings makes the ambiguity explicit rather than forcing a single interpretation that the data cannot support.

4.4 Multilingual Coverage

Of the 856 unique concept sets covered by the mapped Croatian items, 831 (97.1%) contain at least one lexical entry in the nine additional language columns. Coverage is however uneven: German is the best represented with forms for 748 concept sets, followed by Russian (722), French (705), and Spanish (647). Hungarian (515) and Italian (396) occupy a middle position, while Czech (82), Serbian (80), and Ukrainian (8) are sparsely represented. These differences reflect the number and scope of Concepticon lists available for each language rather than any absence of the concepts in those languages. For example, Ukrainian forms appear almost exclusively for colour terms, as the Ukrainian data in Concepticon derive primarily from a colour term study (Jonauskaite et al. 2020).

5 Conclusion and Outlook

This contribution mapped the 1,000 most frequent nouns, verbs, and adjectives from the hrWaC web corpus to existing Concepticon concept sets. After cleaning and translation, the final dataset contains 2,668 items, of which 1,262 (47.3%) were successfully mapped to 856 unique Concepticon concept set identifiers, with lexicalizations from nine additional languages.

The mapping results show that Concepticon covers stable, cross-linguistically recurrent meanings well, but provides weaker coverage for morphologically derived forms and for vocabulary typical of contemporary web-based language use. The

unmapped items are not a failure of the mapping process but a finding in their own right, as they identify specific conceptual areas where Concepticon's current inventory diverges from everyday modern usage. Within the mapped subset, many-to-one mappings are common, and most mapped concepts are also represented in at least one additional language in Concepticon, which confirms that the Croatian items were primarily linked to well-established conceptual cores.

Several limitations of the study are worth acknowledging. The dataset is based on a web corpus, so the frequency distribution reflects genres typical of online discourse, including news, public commentary, and informational texts. Also, spoken or informal vocabulary is likely underrepresented. The mapping relied on English as an intermediate language, which means that language-specific nuances may not always be fully captured by a single English gloss. The manual verification was also carried out by a single annotator, and the dataset would benefit from inter-annotator agreement checks on borderline cases in future work.

As for future directions, the frequency-first workflow used here could be applied to other languages currently absent from Concepticon, enabling broader cross-linguistic comparison of frequency-concept alignment. Methodologically, distributional semantic methods such as word embeddings could support mapping decisions for polysemous and borderline cases where context is unavailable. The full methodology and extended analysis are documented in the underlying Master's thesis (Krišto 2026), to which the reader is referred for further detail.

References

Blum, Frederic and Barrientos, Carlos and Englisch, Janina and Forkel, Robert and Greenhill, Simon and Rzymski, Christoph and List, Johann-Mattis (2025): Lexibank 2: pre-computed features for large-scale lexical data. Open Research Europe 5. 126. https://doi.org/10.12688/openreseurope.20216.2

Dang, Thi Ngoc Yen and Webb, Stuart (2025): Applications of word lists in second language learning and teaching. Language Teaching 58. 1–21.

Forkel, Robert and Rzymski, Christoph and List, Johann-Mattis (2024): PyConcepticon [Python library, Version 3.1.0]. Max Planck Institute for Evolutionary Anthropology: Leipzig. https://pypi.org/project/pyconcepticon

Fulgosi, Sanda and Tuđman Vuković, Nina (2001): Relevantnost frekvencije jezične uporabe pri opisu strukture leksema. Suvremena lingvistika 27. 73–85.

Hudeček, Lana and Mihaljević, Milica (2019): Croatian web dictionary – Mrežnik vs. Croatian linguistic terminology – Jena. In: Proceedings of the 7th International Conference The Future of Information Sciences INFuture2019: Knowledge in the Digital Age. 22–31. https://doi.org/10.17234/INFUTURE.2019.4

Jonauskaite, Domicele and Parraga, C. Alejandro and Quiblier, Michael and Mohr, Christine (2020): Feeling blue or seeing red? Similar patterns of emotion associations with colour patches and colour terms. Perception 11. https://doi.org/10.1177/2041669520902484

Kapović, Mate (2022): Strani jezici u formalnom obrazovanju u Hrvatskoj. Strani jezici 51. 283–309.

Kilgarriff, Adam (2010): Comparable corpora within and across languages, word frequency lists and the KELLY project. Lexical Computing Ltd, Brighton.

Krišto, Anja and Mambrini, Francesco (2026): Integrating Croatian into Concepticon: a corpus-based frequency mapping of Croatian vocabulary. Master's thesis, Università Cattolica del Sacro Cuore, Milan. https://doi.org/10.17613/p4fqh-03m92

List, Johann-Mattis and Cysouw, Michael and Forkel, Robert (2016): Concepticon: A resource for the linking of concept lists. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2393–2400. https://aclanthology.org/L16-1379

List, J.-M., A. Tjuka, F. Blum, A. Kučerová, C. Barrientos Ugarte, C. Rzymski, S. Greenhill, and R. Forkel (2025): CLLD Concepticon [Dataset, Version 3.4.0]. Max Planck Institute for Evolutionary Anthropology: Leipzig. https://concepticon.clld.org

List, Johann-Mattis and Forkel, Robert and Greenhill, Simon J. and Rzymski, Christoph and Englisch, Johannes and Gray, Russell D. (2022): Lexibank: A public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9. 44. https://doi.org/10.1038/s41597-022-01432-0

Ljubešić, Nikola and Erjavec, Tomaž (2011): hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In: Proceedings of the 14th International Conference on Text, Speech and Dialogue. 395–402. https://doi.org/10.1007/978-3-642-23538-2_50

Nation, I.S.P. (2001): Learning Vocabulary in Another Language. Cambridge University Press, Cambridge.

Peti-Stantić, Anita and Anđel, Maja and Gnjidić, Vedrana and Keresteš, Gordana and Ljubešić, Nikola and Masnikosa, Irina and Tonković, Mia and Tušek, Jana and Willer-Gold, Jana and Stanojević, Mateusz-Milan (2021): The Croatian psycholinguistic database: Estimates for 6000 nouns, verbs, adjectives and adverbs. Behavior Research Methods 53. 1799–1816. https://doi.org/10.3758/s13428-020-01533-x

Tjuka, Annika and Forkel, Robert and List, Johann-Mattis (2023): Curating and extending data for language comparison in Concepticon and NoRaRe. Open Research Europe 2. 1–11. https://doi.org/10.12688/openreseurope.15380.3

Supplementary Material
Data and code can be found at https://github.com/anjak00/croatian-frequency-wordlist-concepticon
Acknowledgements
I would like to thank Francesco Mambrini for his supervision of the Master's thesis underlying this contribution, and Johann-Mattis List for his guidance in preparing this work for publication.