Exploring Derivational Families through Intelligent Lexicography

Krešimir Šojat[1], Kristina Kocijan[2]

[1] Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3 Zagreb Croatia

[2] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3 Zagreb Croatia

e-mail: ksojat@ffzg.unizg.hr, krkocijan@ffzg.unizg.hr

Abstract

This paper presents a novel approach to exploring derivational families within the framework of Intelligent Lexicography, using the ŠKOLARAC corpus: a collection of Croatian school essays written by L1 learners (native-speaking students) in grades 5 through 8 and enriched with metadata such as gender, grade level, and region. By combining rule-based linguistic processing in NooJ, a linguistic development environment for formalizing morphological and syntactic patterns, with tailored morphological procedures for Croatian, the study identifies and maps derivational networks of three pedagogically relevant lexical morphemes (CRT, PIS, and RAD) tracing their associated inflected and derived forms as they appear in young learner corpora. The extracted data are visualized using radial graphs, butterfly charts, and hierarchical structures, enabling a multifaceted analysis of morphological productivity and lexical variation. This integrated workflow demonstrates how intelligent tools can enhance lexicographic practice by uncovering deep morphological relationships in authentic learner language. The findings support the development of adaptive, learner-sensitive lexicographic resources with applications in linguistics, language education, and curriculum design, particularly in the context of developing digital dictionaries and vocabulary tools tailored to young learners.

Keywords: intelligent lexicography; derivational families; learner corpora; Croatian morphology; linguistic visualization; ŠKOLARAC corpus

1. Introduction

In recent years, lexicography has undergone a significant transformation, driven by advances in computational linguistics and digital technologies. At the forefront of this evolution is Intelligent Lexicography, which integrates linguistic theory with natural language processing to offer innovative methods for analyzing, representing, and interpreting lexical data. This fusion is particularly valuable for exploring complex morphological phenomena such as derivational word families, where understanding the relationships between roots (lexical morphemes), stems and derivatives sheds light on both language structure and usage patterns. As Lew (2024) points out, the role of lexicographers is not being diminished but redefined in the age of AI. This study

contributes to that redefinition by demonstrating how linguistically-informed, rule-based approaches, especially when applied to underrepresented languages and learner data, can serve as a form of intelligent lexicography that complements data-driven and AI-based methods.

Croatian, as a South Slavic language, exhibits a rich inflectional and word-formational morphology. The two primary word-formation processes in Croatian are derivation and compounding. The main difference between them lies in the number of lexical morphemes involved: derivation operates on a single lexical morpheme, while compounding combines two or more. Both processes typically involve affixation. The affixes used in Croatian include prefixes and suffixes in both derivation and compounding, as well as interfixes in compounding. Inflection in Croatian relies exclusively on suffixation, whereas word-formation mechanisms include suffixation, prefixation, simultaneous use of both, and ablaut. A derivational family consists of words that share the same lexical morpheme but differ according to the affixes used in their formation. This study focuses exclusively on derivation, leaving compounding for future research.

Derivational morphology plays an important role in shaping lexical richness, supporting both linguistic creativity and vocabulary development. In morphologically rich languages such as Croatian, the structure of derivational families provides insight into how word-formation processes reflect patterns of meaning, learning, and stylistic variation. However, most existing studies rely on standardized corpora, often overlooking the nuanced and unpredictable behaviors found in learner generated texts. This is particularly relevant for lexicography, which increasingly seeks to incorporate authentic, user-based data into dictionary design, especially in pedagogical contexts (Gabrielatos, 2005; Frankenberg-Garcia, 2014).

The ŠKOLARAC corpus provides a rare opportunity to examine derivational phenomena in authentic writing by Croatian primary school students. Comprising essays written both by female and male students from grades 5 through 8, and enriched with demographic metadata (gender, grade level, region), the corpus enables a fine-grained exploration of how young learners engage with derivational morphology in real-world contexts. Additionally, it supports the investigation of potential gender-based variation in morphological awareness and lexical productivity. The fact that these are L1 learners further strengthens the educational relevance of the study, offering insight into natural vocabulary development rather than second-language acquisition.

The aim of this study is to analyze how members of derivational families are distributed and activated across the two learner groups. A prerequisite for this type of research is the identification and processing of words that share the same lexical morpheme, that is, the compilation and modeling of derivational families. By identifying and mapping root forms within each subcorpus, and visualizing their associated networks and structural patterns, we reveal both shared and divergent tendencies in word formation.

In doing so, we contribute to the fields of corpus-based morphology and intelligent lexicography, highlighting the potential for learner-informed lexicographic tools that reflect real usage patterns and support vocabulary growth in educational settings.

Our methodology combines rule-based linguistic annotation in NooJ (a platform for formalizing morphological, syntactic, and semantic rules) with custom computational procedures designed to identify and group derivationally related words. By anchoring each token to its root and aggregating frequency data across inflectional forms, we enable detailed analysis of the use of specific members within derivational families, identification of the most frequently used family members and observation of underlying morphological patterns. These results are subsequently visualized through interactive tools such as Tableau and Voyant Tools, allowing for multifaceted exploration of derivational structures and their distribution across demographic subgroups.

Through this study, we demonstrate how Intelligent Lexicography can effectively combine computational methods, linguistic expertise, and visualization technologies to deepen our understanding of word formation in educational contexts. This aligns with the broader aims of the eLex conference, showcasing the potential of intelligent, data-driven lexicography in the 21st century.

Building on this conceptual foundation, the remainder of the paper is structured as follows. Section 2 provides an overview of related research in derivational morphology, learner corpora, and intelligent lexicography, positioning our study within current scholarly discourse. Section 3 introduces the ŠKOLARAC corpus and presents key quantitative and linguistic characteristics of its gender-based subcomponents. Section 4 outlines the methodological workflow, from transcription through morphological processing and derivational tagging, detailing the role of rule-based parsing and visualization tools. Section 5 presents case studies of three selected root families, CRT, PIS, and RAD, highlighting gendered patterns of derivational usage in authentic student writing. Section 6 discusses the educational implications of our findings, with a focus on learner-sensitive lexicographic applications. Finally, Section 7 summarizes the main contributions of the study and proposes directions for future work.

2. Background and related work

This study lies at the intersection of three research strands: derivational morphology, learner corpora, and intelligent lexicography. We briefly review each and then position our work in relation to existing resources and methods.

2.1 Derivational Morphology in Croatian

Derivational morphology examines how roots combine with affixes to form new words. The analysis of derivational morphology has long been central to morphological theory

and development of lexical resources, especially in languages with rich inflectional and derivational systems such as Croatian (Vučković et al., 2010; Kocijan et al., 2018). Traditional approaches have focused on identifying affixes, their productivity, and their use in the formation of specific parts of speech (Babić, 2002; Barić et al., 2003; Silić & Pranjković, 2005). However, while numerous studies have addressed derivational morphology (Hržica, 2021; Kocijan, 2022; Kuna, 2022), few have explored it in the context of authentic learner-produced texts with detailed demographic metadata, particularly in a morphologically rich language like Croatian.

Derivational processes in Croatian contribute substantially to lexical expansion and semantic nuance (Šojat et al., 2012). The development of derivational resources such as CroDeriv (Šojat et al., 2014; Filko et al., 2020; Šojat & Filko, 2023) has significantly advanced the understanding of derivational families and processes by systematically mapping relationships between lemmas and their derivatives. NLP tools like NooJ (Silberztein, 2016) have supported the formalization of grammatical descriptions and enabled the automation of morphological analysis, both inflectional and derivational (Tadić & Fulgosi, 2003; Vučković et al., 2011; Vučković et al., 2013; Kocijan et al., 2016; Šojat et al., 2019; Kocijan & Šojat, 2024). Despite these advances, the application of such tools to learner language, particularly in educational settings, remains relatively underexplored. This gap highlights the importance of studies that integrate detailed morphological analysis with pedagogically relevant corpora, such as ŠKOLARAC. By pioneering the use of these computational and linguistic technologies on the ŠKOLARAC corpus, this study offers novel insights into the functioning of derivational morphology in authentic learner language and educational contexts.

2.2 Learner Corpora and Morphological Awareness

Sociolinguistic research consistently highlights gender-based differences in language production, including vocabulary size, stylistic preferences, and lexical creativity (Llach, 2010; Teibowei, 2024). Within learner corpora, these differences can illuminate how male and female students diverge in morphological choices, thematic focus, and error patterns. Despite this potential, few studies have compared derivational productivity across gender in primary school writing.

Learner-produced texts introduce non-standard forms and orthographic variation that are typically absent from edited corpora. Studies of learner language have shown that spelling errors, emergent derivations, and irregular affix use reflect learners' evolving understanding of lexical structure, e.g. progressing from root recognition toward productive affix application. These features are especially pronounced in early educational contexts, where lexical experimentation forms part of linguistic development. Still, most work to date has treated such phenomena as noise rather than as data points that can serve as valuable indicators of morphological development and acquisition stages (Murakami & Alexopoulou, 2016).

The ŠKOLARAC corpus, comprising over 1.8 million tokens of school essays written by primary school students (grades 5 - 8), has a potential to offer a unique window into the extent to which young learners use derivational patterns in Croatian, as well as the scope of vocabulary use within individual derivational families in authentic writing. Its gender-annotated subcorpora allow for comparative exploration of how children engage with word formation processes, not only in terms of frequency and diversity, but also in relation to pedagogical and stylistic factors. The presence of both standardized and non-standard word forms enables a dual perspective, capturing both normative morphology and learner innovation.

2.3 Intelligent Lexicography

Intelligent Lexicography extends beyond traditional dictionary compilation by integrating computational linguistics, corpus data, and interactive visualisations. Recent research emphasizes the need for adaptive, pedagogically informed linguistic resources that respond to authentic usage data and support language acquisition (Quixal et al., 2021; Römer-Barron, 2024). These approaches combine computational methods and lexical data to support practical applications, such as educational tools, learner diagnostics, and curriculum-sensitive language materials.

While recent developments in intelligent lexicography often focus on the integration of large language models (LLMs), our approach demonstrates that rule-based tools such as NooJ, combined with corpus data and visualization platforms like Tableau and Voyant Tools, offer robust and transparent methods for exploring complex morphological phenomena, particularly in learner language where precision and interpretability are crucial. This distinction is important as our study showcases intelligent lexicographic practices that do not rely on LLMs but still contribute significantly to the field through the use of linguistic expertise and computational resources.

Previous work has demonstrated the pedagogical and analytical value of visualizing linguistic resources using platforms such as Tableau and Cytoscape[1], particularly in the context of Croatian morphology (Kocijan, 2015). Such visualizations facilitate the interpretation of complex morphological structures and support learner-sensitive exploration of lexical relationships.

At its core, intelligent lexicography holds great potential to leverage real-world language data for language acquisition research, textbook design, and even second-language instruction. This potential can be realized through rule-based parsers that automate morphological annotation, custom algorithms for root identification and derivational mapping tailored to a language's morphology, and visualization techniques, such as

[1] Originally developed for visualizing molecular interaction networks in bioinformatics and chemistry, Cytoscape has proven adaptable for representing complex linguistic structures, such as morphological relationships in Croatian (Kocijan, 2015).

network graphs, radial trees, and butterfly charts, that reveal complex lexical relationships and usage patterns. Together, these methods offer powerful potential to provide great insights in the analysis of learner corpora and enable dynamic exploration of derivational families.

Our work addresses three interrelated gaps: first, we move beyond edited corpora by analysing authentic student writing to enable learner-focused derivational analysis; second, we integrate NooJ's parsing capabilities with custom scripts and interactive visualizations (Tableau and Voyant Tools) to map derivational families end-to-end; and third, by separating the ŠKOLARAC corpus into girls' and boys' subcorpora, we reveal how these families are differentially activated in both form and frequency. In so doing, we demonstrate a practical application of intelligent lexicography that is responsive to learner data, morphologically informed, and pedagogically relevant.

3. Corpus description

The ŠKOLARAC corpus is a collection of authentic texts written by Croatian elementary school students, created with the aim of supporting linguistic, lexicographic, and pedagogical research. It consists of manually collected and curated essays written by pupils from grades 5 to 8, originally submitted either as handwritten paper documents or scanned images sent via email. In both cases, the texts were handwritten. The task of manual transcription was to enrich the text with metadata and record the text in its original form, preserving all errors made by the students without correcting any mistakes that may have been amended by the teacher. For the purpose of this study, we focus exclusively on two gender-based subsets: texts written by girls and those written by boys. Both subcorpora were linguistically annotated in NooJ and prepared for morphological analysis. The key differences in size and lexical richness are presented in two tables (Table 1 and Table 2).

Subcorpus	**Texts**	**Tokens (incl. digits)**	**Unique Tokens**	**Digit Tokens**
GIRLS	2,105	1,325,892	47,025	19,189
BOYS	1,957	523,651	43,319	19,058

Table 1. Corpus ŠKOLARAC statistics

A comparison of the subcorpora reveals notable differences in both volume and lexical variety. The girls' subcorpus contains slightly greater number of texts (2,105 vs. 1,957), yielding over 1.3 million tokens versus approximately 524,000 found in boys' subcorpus. Girls' texts also exhibit greater lexical variety (47 025 vs. 43 319 unique tokens), reflecting richer vocabulary use and higher orthographic variation (including misspellings and creative forms).

The observed imbalance between the girls' and boys' subcorpora in terms of tokens and

texts reflects the natural distribution of available data rather than an artificially balanced sample. This choice was deliberate to preserve representativeness and ecological validity, ensuring the analysis reflects genuine learner behavior in authentic educational contexts. This means we do not artificially balance the subcorpora by removing texts from girls, as doing so would reduce the naturalness and diversity of the dataset and potentially obscure real usage patterns. This approach enables us to investigate gender-specific morphological awareness and lexical productivity patterns in detail, thereby contributing to the sociolinguistic and educational understanding of language acquisition processes.

Despite the imbalance in total tokens, the lemmatized vocabulary shows identical counts of basic forms for adjectives, nouns, and verbs (9,987 in each subcorpus). However, there are subtle internal differences: boys' essays contain slightly higher proportion of adjectives, while girls' texts include more nouns and verbs. These distributional tendencies may reflect stylistic or topic preferences in their writing.

Subcorpus	Total Lemmas (Adj/Noun/Verb)	Adjectives	Nouns	Verbs
GIRLS	9,987	2,327	**5,386**	**2,274**
BOYS	9,987	**2,467**	5,381	2,139

Table 2. Lexical Categories by Lemma (Adjectives, Nouns, Verbs only)

Both subcorpora were processed using the NooJ linguistic resources, where morphological parsing and lexical annotation were conducted. Prior to analysis, we verified that all forms related to the roots CRT, PIS, or RAD were already covered in the NooJ dictionary, and no new variants specific to these roots appeared in the corpus. For other roots, this verification remains to be performed in further stages. This ensured that the inflectional and derivational tagging related to these root families was accurate and comprehensive.

Building on these corpus-based observations, the following section outlines the methodological framework used to extract and visualize derivational relations within the ŠKOLARAC corpus. We describe the combination of rule-based parsing, recognition of derivational patterns, and custom algorithmic procedures, followed by the integration of these outputs into visual models that allow for multifaceted exploration of derivational networks.

4. Methodology: from transcription to derivational analysis

The analytical workflow begins with manual transcription of all handwritten ŠKOLARAC essays into digital text, carefully preserving original orthography, including spelling errors and punctuation, and adding metadata (authors' gender, grade

level, region) for each text file. For the purposes of this study, only the gender variable was utilized, as other metadata are still being processed.

Once digitized, the essays were loaded into the NooJ environment and organized into two gender-specific subcorpora (girls vs. boys), ensuring that every subsequent processing step could be carried out separately for each group. The analytical process that started with manual transcription, culminated in a derivational analysis focused on three productive root families.

Using Croatian linguistic resources available in NooJ, each corpus was processed to assign basic part-of-speech (POS) tags to all tokens. The built-in morphological dictionaries and grammars provided lemmatization and POS annotation for standard word forms. A list of unknown words was automatically generated during this process. This list was manually reviewed and correctly spelled, but previously unrecognized words, were added to the NooJ dictionary to improve parsing coverage. Importantly, misspelled words were retained unaltered, as they represent valuable indicators of learner strategies and acquisition challenges, offering rich data for future studies on learner errors and orthographic variation.

For derivational analysis, we selected three early-acquired and pedagogically salient root families, CRT ('to draw'), PIS ('to write'), and RAD ('to work'), on the basis that their base forms appear in learners' vocabulary from preschool onward. In the NooJ dictionary, all lemmas derived from these roots were flagged with a custom attribute (Root=CRT, Root=PIS, Root=RAD). Compounds and blends containing these roots in combinations with other stems (such as *pismonoša* 'postman' or *radoholičar* 'workoholic') were deliberately excluded and set aside for future study.

To extract every inflected and derived form of the tagged lemmas, we designed a custom syntactic grammar in NooJ. This grammar scanned each corpus for tokens marked with our root attributes, annotating each occurrence with its token, i.e. original word form as found in the text, its corresponding lemma, and assigned root value. Applying this grammar across both subcorpora yielded comprehensive lists of relevant word forms and derivatives.

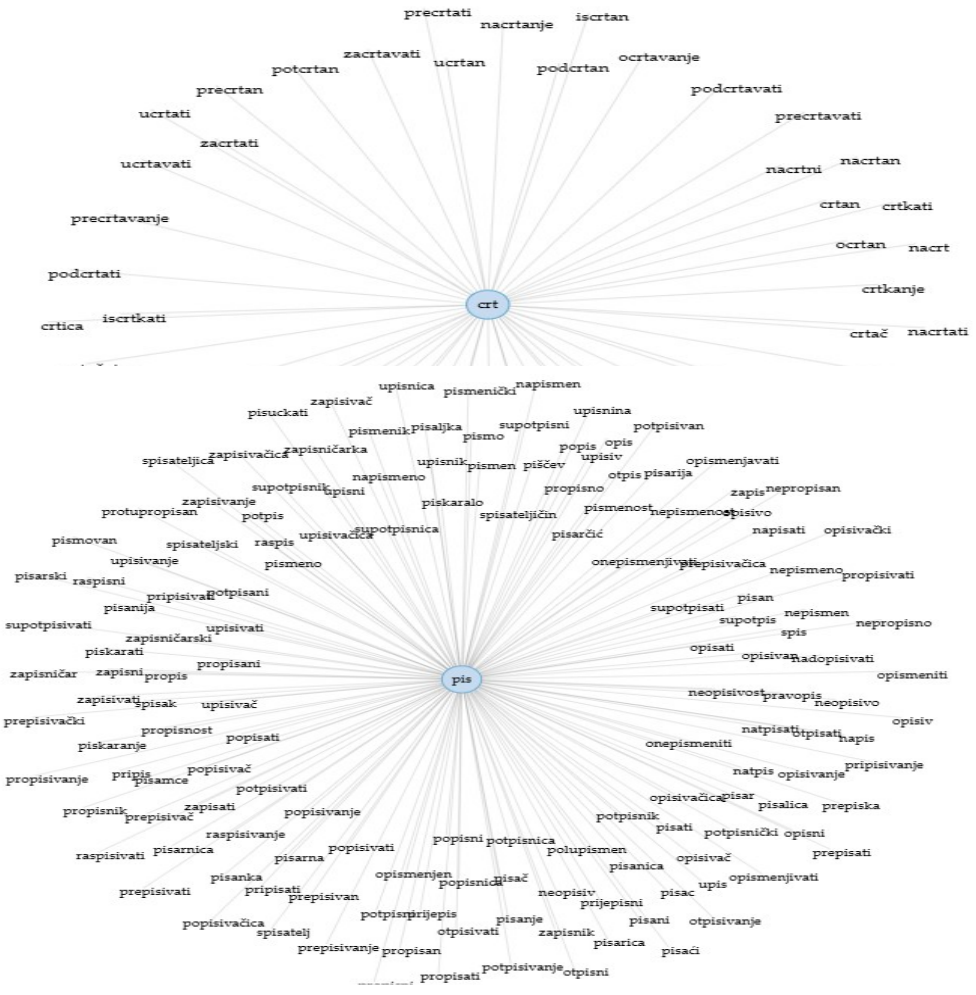
Finally, the annotated outputs were exported from NooJ as plain-text files, then loaded into Excel for cleaning, aggregation, and preliminary tabulation. The cleaned datasets were imported into Tableau, where we constructed interactive visualizations, most notably butterfly charts comparing lemma usage by gender, and radial and hierarchical graphs depicting the internal structure of each derivational family. These visual tools support a multifaceted exploration of morphological productivity, lexical variation, and gender-based patterns in authentic learner writing.

5. Case studies: derivation families of CRT | PIS | RAD

In this section, we present a multifaceted exploration of three derivational families built around the roots CRT (‘to draw’), PIS (‘to write’), and RAD (‘to work’) in the ŠKOLARAC corpus. We combine radial visualizations, bar-chart comparisons, proportional measures, and statistical tests to reveal how these roots are activated and which derivatives are used by male and female students in authentic learner texts.

5.1 Radial Visualization of Derivational Productivity

Figures 1, 2, and 3 arrange all lemmas derived from each root in a radial layout, with the root positioned at the center and linked to its derivatives. This immediately illustrates the derivational breadth of each family: CRT, comprising 52 lemmas, appears less sprawling than the more productive PIS (177 lemmas) and RAD (110 lemmas) families, yet even CRT supports a rich network of nouns, verbs, and adjectives. The derivational families used in this study were extracted from the CroDeriv database (Šojat & Filko, 2023) which systematically maps morphological relationships among Croatian lexemes.



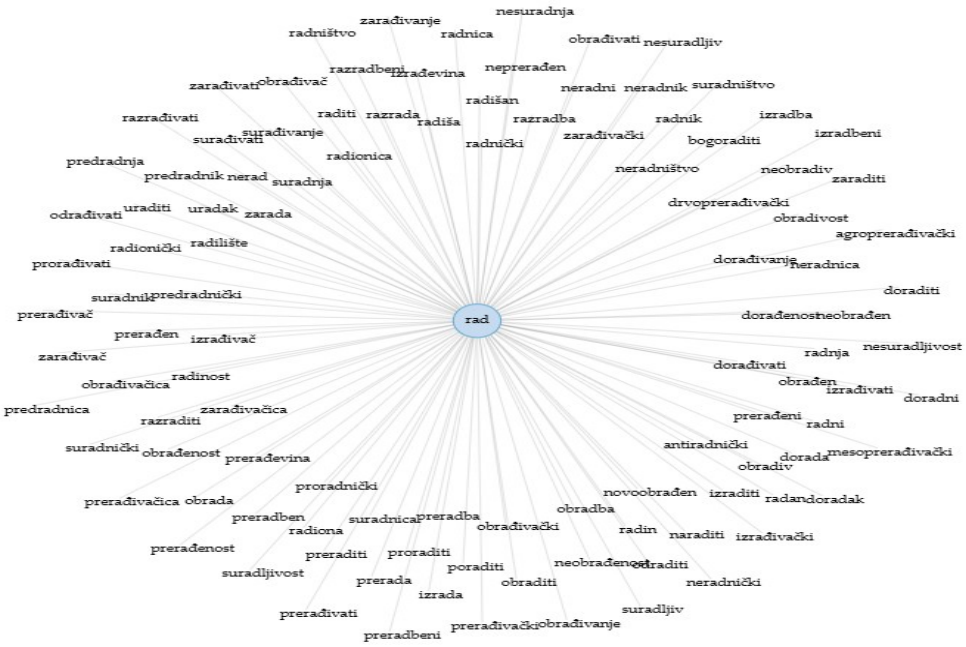


Figure 3. Radial derivational graphs of all lemmas derived from roots RAD as listed in the NooJ dictionary.

5.2 Distinct Lemmas and Tokens by Gender

Figure 4 compares the number of distinct lemmas each gender uses per root. Girls employ slightly more derivational types across the board (CRT: 21 vs. 15; PIS: 42 vs. 40; RAD: 37 vs. 30), suggesting a broader lexical range in their writing.

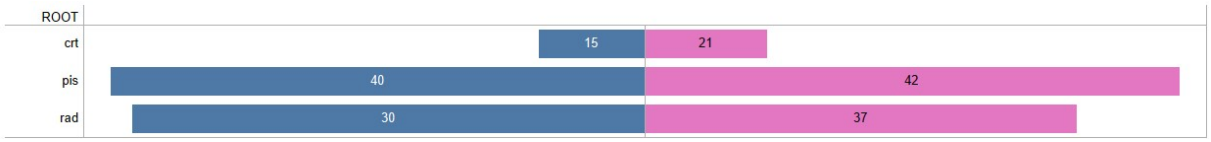


Figure 4. Number of distinct lemmas derived from the roots CRT, PIS, and RAD used by boys and girls in the ŠKOLARAC corpus.

When we shift from types to tokens (Figure 5), the gap widens. Girls not only use more forms but also repeat them more frequently (CRT: 368 vs. 246; PIS: 1 611 vs. 1 248; RAD: 1 303 vs. 1 196). These raw counts reflect both the larger size of the girls' subcorpus (1.3 million vs. 524 thousand tokens) and a genuine tendency toward greater engagement with these derivational families.

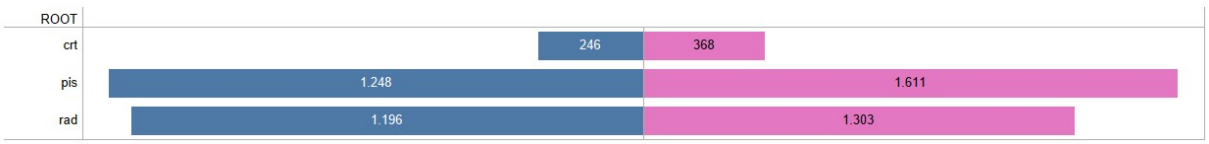


Figure 5. Total number of word occurrences (tokens) derived from the roots CRT, PIS, and RAD used by boys and girls in the ŠKOLARAC corpus.

When interpreting these findings, it is important to consider the overall size of each

subcorpus. The girls' subcorpus contains slightly more texts (148 more than boys') but significantly more tokens (1.3 million vs. 524,000). This naturally contributes to higher absolute frequencies. However, even when accounting for this difference, the relative proportions still suggest that girls make more varied use of derivational forms tied to these roots.

To gain a more detailed view of how derivational families are used across genders, we visualized the distribution of individual lemmas derived from the roots CRT, PIS, and RAD using butterfly charts (Figures 6–8). Each chart displays the number of distinct lemmas used by boys (left, blue bars) and girls (right, pink bars), sorted by frequency in the girls' subcorpus.

By sorting the charts according to the frequency of use among girls, the visual structure emphasizes the lexical richness and expressive tendencies in their writing. In contrast, boys' usage is often concentrated around a smaller set of high-frequency lemmas, suggesting a more focused or utilitarian lexical strategy.

These charts not only illustrate gender-based variation in derivational usage but also highlight the pedagogical potential of such analyses, for example, in identifying underused lexical items, supporting vocabulary expansion, or tailoring language instruction to learner profiles.

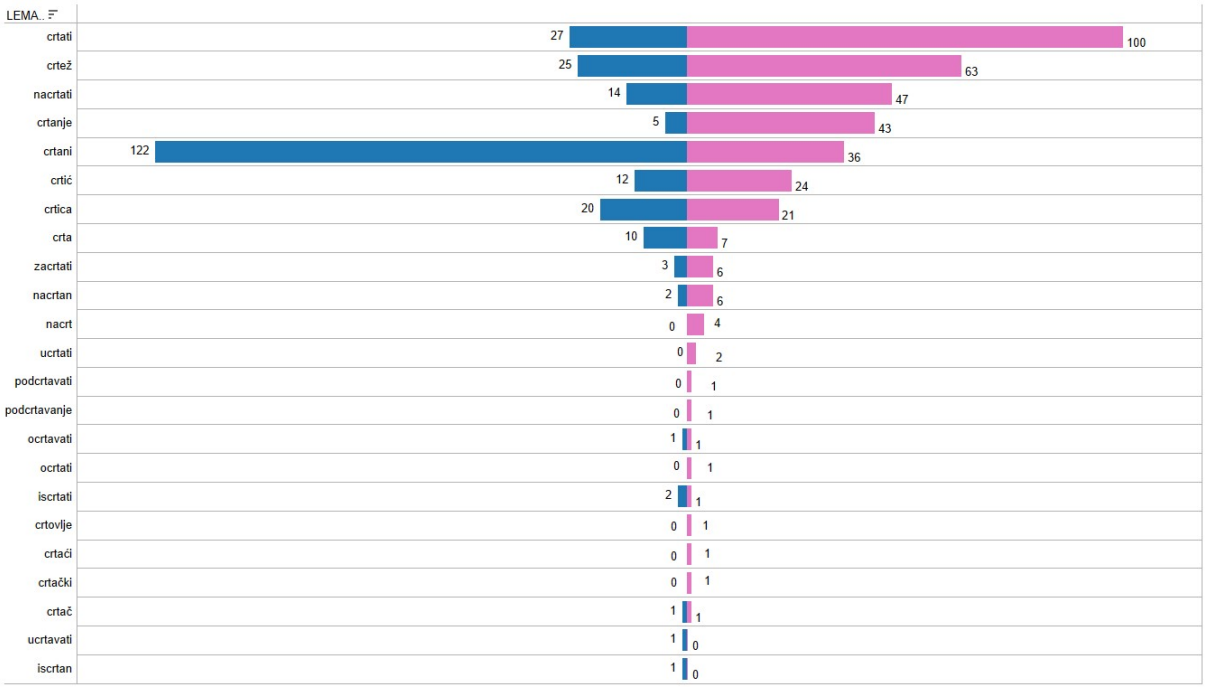


Figure 6. Butterfly chart showing the number of distinct CRT-root lemmas used by boys (left, blue) and girls (right, pink) in the ŠKOLARAC corpus. Lemmas are sorted by frequency in the girls' subcorpus.

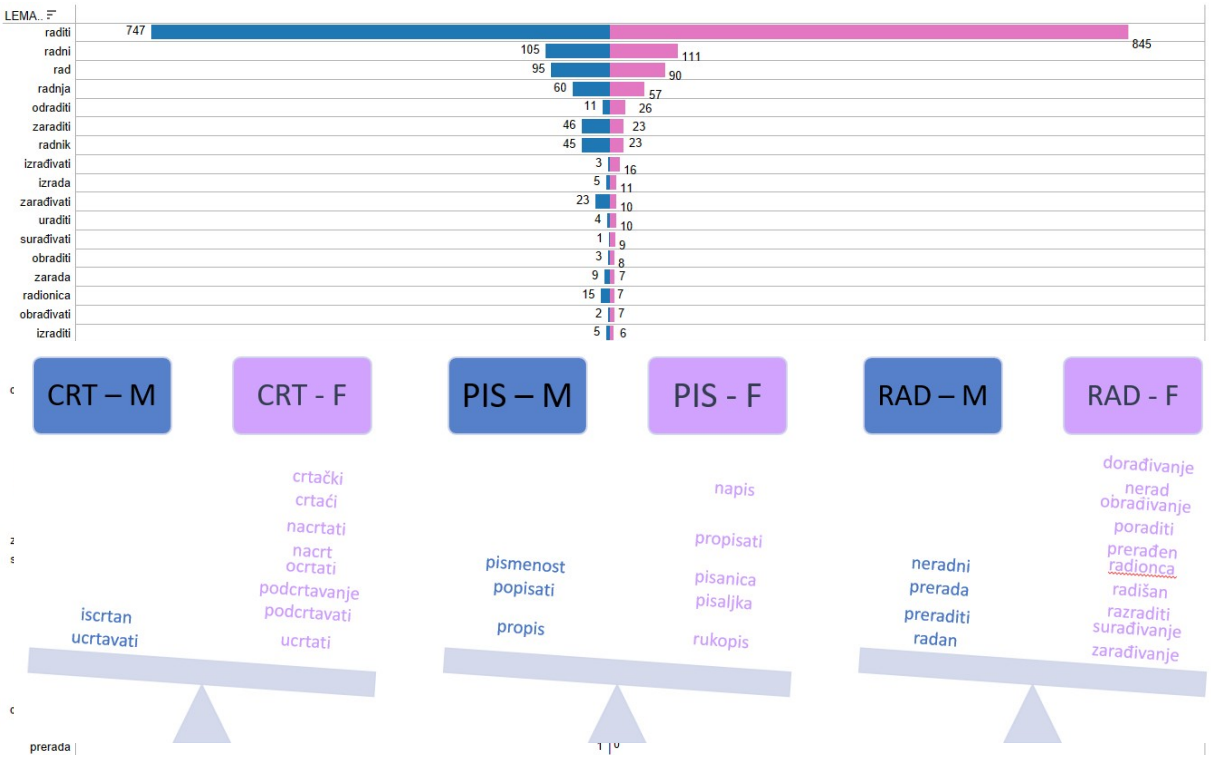


Figure 9. Seesaw diagram showing gender-exclusive lemmas derived from the roots CRT, PIS, and RAD in the ŠKOLARAC corpus. Each side of the seesaw represents lemmas used exclusively by boys (left) or girls (right), with the tilt reflecting the lexical imbalance.

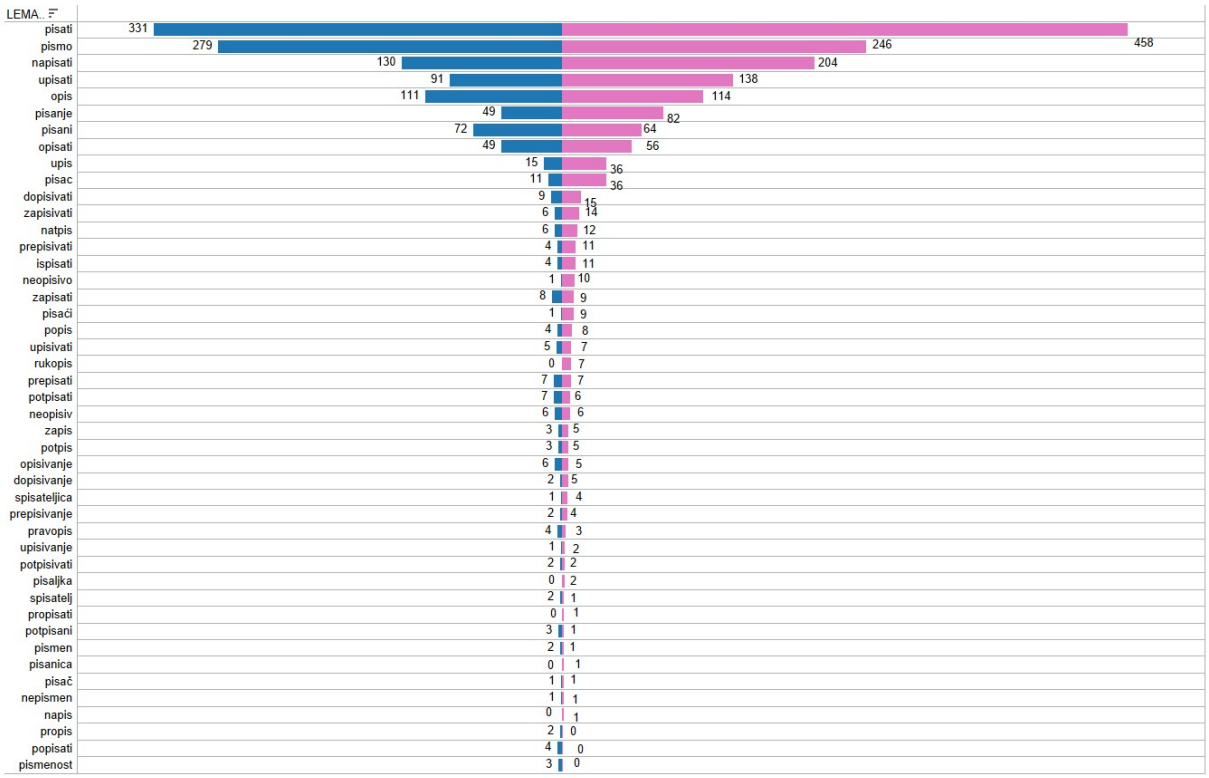


Figure 8. Butterfly chart showing the number of distinct PIS-root lemmas used by boys (left, blue) and girls (right, pink). Lemmas are sorted by frequency in the girls' subcorpus.

To examine lexical divergence within derivational families, we extracted all lemmas

derived from the roots CRT, PIS, and RAD that appeared exclusively in either the boys' or girls' subcorpus. These gender-exclusive lemmas are visualized in the form of seesaw diagrams (Figure 9), where each side represents the number and type of lemmas unique to one group. The diagrams are intentionally tilted toward the girls' side, reflecting the greater lexical variety observed in their writing.

The data (Figure 9) reveal clear asymmetries:

- CRT root: Boys used 2 exclusive lemmas (*iscrtan* 'drawn', *ucrtavati* 'to draw in'), while girls used 9, including *nacrtati* ' to draw' *nacrt* ' draft, design', *crtaći* 'drawing', and *podcrtavanje* ' underlining'. These forms suggest a broader engagement with both verbal and nominal derivations, particularly in creative and descriptive contexts.

- PIS root: Boys contributed 3 exclusive lemmas (*propis* 'regulation', *pismenost* 'literacy', *popisati* 'to list, to inventory'), while girls used 5, such as *pisaljka* 'writing utensil', and *napis* 'sign, label'. Girls' forms tend to reflect expressive or personal writing, while boys' forms lean toward formal or administrative vocabulary.

- RAD root: Boys used 4 exclusive lemmas (*preraditi* 'to revise, to rework', *radni* 'working, operational', *neradni* 'non-working', *prerada* 'processing, modification'), whereas girls used 9, including *surađivanje* ' cooperation', *zarađivanje* ' earning', *radišan* 'hardworking, diligent' and *doređivanje* 'refinment, adjusment'. Girls' lemmas often involve collaborative, affective, or nuanced verbal forms, suggesting a wider semantic and functional range.

These patterns reinforce earlier findings that girls tend to engage more deeply with derivational families, both in terms of frequency and lexical creativity. The presence of gender-exclusive lemmas also suggests that certain derivational forms may be more closely tied to stylistic preferences, narrative strategies, or topic selection that differ between boys and girls. By highlighting not just what is shared, but what is absent in each group, these diagrams offer a nuanced view of how derivational morphology reflects broader patterns of language use in learner writing.

5.3 Proportional Coverage, Normalized Frequency and Log-Likelihood

Analysis

To quantify these observations, we computed three measures. The proportional lemma coverage (# of used lemmas / # of total lemmas in dictionary for specific root x 100) reveals that girls consistently activate a larger portion of the available derivational lexicon for each root. For example, girls use 40.4% of all CRT-root lemmas listed in the dictionary, compared to 28.8% for boys. This pattern holds across all three roots (PIS:

F-23,7% vs M-22,6%; RAD: 33,6% vs M-27,3%), suggesting that girls engage more broadly with derivational families, both in terms of lexical variety and morphological depth.

However, when normalized per 10,000 tokens (root token count / total tokens in subscorpus x 10 000), boys actually use CRT-, PIS-, and RAD-root words more frequently than girls, despite having fewer total tokens. For instance, boys use RAD-root words at a rate of 22.83 per 10,000 tokens, compared to 9.83 for girls (PIS: F-12,15% vs M-23,83%; CRT: F-2,77% vs M-4,70%). This suggests that while girls use a wider variety of derived forms, boys tend to repeat a smaller set of derivational forms more intensively.

To assess whether certain lemmas are used significantly more by one gender, we conducted a log-likelihood (LL) analysis comparing their frequencies in the boys' and girls' subcorpora. The results reveal several statistically significant differences:

- The verb ***crtati*** 'to draw' is used significantly more by girls (LL = 20.63, $p < 0.001$), while the adjective *crtani* 'drawn, animated' is strongly associated with boys (LL = 63.91, $p < 0.001$). This suggests a stylistic divergence: girls may focus more on the act of drawing, while boys refer more often to *crtani film* 'animated film', a common collocation in their texts.

- The noun ***upis*** 'enrollment, registration' also shows a significant gender difference (LL = 4.94, $p < 0.05$), appearing more frequently in girls' writing.

- The verb ***raditi*** 'to work' is highly frequent in both subcorpora, but the LL score (14.45, $p < 0.001$) indicates a relatively stronger association with boys, despite girls having a larger corpus overall.

- The noun ***radnik*** 'worker' is significantly more frequent in boys' texts (LL = 7.14, $p < 0.01$), pointing to a more concrete use of the RAD root in male-authored essays.

Interestingly, the lemma ***pisati*** 'to write', despite being the most frequent among the PIS-root forms, does not show a statistically significant difference (LL = 0.12), indicating balanced usage across genders.

These findings underscore the value of LL analysis in identifying lexical preferences and stylistic tendencies that may not be visible through raw frequency alone. They also reinforce the broader pattern observed in the corpus: girls tend to use a wider range of derived forms, while boys often concentrate on a smaller set of high-frequency, semantically focused lemmas.

Together, these case studies demonstrate how combining visual, proportional, and statistical analyses can uncover both shared foundations and gender-specific patterns in derivational morphology. The interplay of breadth (types), depth (tokens), and

significance (LL) offers a comprehensive picture of how young learners activate word-formation processes in real writing.

6. Educational implications

The observed differences in the use of derivational families between boys and girls in the ŠKOLARAC corpus carry important implications for language education and curriculum design. While both groups demonstrate active engagement with morphologically rich vocabulary, the analysis of gender-exclusive lemmas reveals qualitative distinctions in lexical development and stylistic expression.

Girls consistently used a wider range of derived forms, particularly those associated with affective, creative, or process-oriented meanings (e.g., *nacrtati* 'to draw', *zarađivanje* 'earning', *doređivanje* 'adjustment, finetuning'). These forms suggest a tendency toward narrative elaboration, emotional nuance, and collaborative or introspective themes. In contrast, boys' exclusive lemmas were fewer and more concrete, functional, or nominal, indicating a more task-oriented or factual style of expression.

From an educational perspective, these findings highlight the need to:

- encourage morphological awareness: teaching strategies can explicitly draw attention to derivational families and their semantic potential, helping learners expand their vocabulary through pattern recognition and analogy;
- diversify writing prompts: by offering a broader range of genres and topics, especially those that invite emotional, descriptive, or imaginative language, teachers can support more balanced lexical development across genders;
- tailor vocabulary instruction: recognizing which derivational forms are underused by each group allows educators to design targeted interventions; for example, boys may benefit from activities that promote expressive or affective vocabulary, while girls might be encouraged to explore more technical or procedural terms;
- integrate corpus-informed materials: the inclusion of real learner data in teaching materials, such as examples of derivational families used in authentic student writing, can make morphological instruction more relevant and engaging.

Ultimately, the presence of gender-exclusive lemmas underscores the importance of lexical diversity as a developmental goal, not only in terms of quantity but also in terms of semantic range and stylistic flexibility. Derivational analysis thus offers a valuable lens for identifying lexical gaps, pedagogical opportunities, and learner-specific needs in the classroom.

The gender-based differences in derivational usage identified in the ŠKOLARAC corpus

offer valuable guidance for the design of prospective digital lexicographic tools tailored to young language users. These envisioned tools could move beyond static dictionary entries by incorporating adaptive, learner-sensitive features that reflect real usage patterns and foster vocabulary growth through enhanced morphological awareness. Building on these insights, we outline a set of features that lexicographic tools could incorporate to support vocabulary development and foster morphological awareness among young learners:

1. ***Morphological Family Navigation*** could enable learners to visually explore derivational families (e.g., radial graphs or tree structures), highlight frequently used forms in learner corpora and suggest underused but related forms to encourage lexical expansion. Thus, when a student looks up *pisati* 'to write', the tool could show *pisaljka* 'writing utensil, stylus'*,* *napis* 'inscription, sign, label', *prepisivati* 'to copy, to cheat on the test', with usage examples drawn from authentic student texts.

2. ***Personalized Vocabulary Suggestions*** could be generated using corpus-based insights to identify lexical gaps associated with specific learner profiles (e.g., boys underusing affective RAD-forms). By suggesting derivationally related words that are less frequent in the learner's demographic group can promote balanced lexical development.

3. ***Contextualized Examples from Peer Writing*** might draw directly from the ŠKOLARAC corpus, showcasing how peers use derived forms in real contexts and also allow learners to compare how the same root (e.g., CRT) is used differently across genres or by different groups.

4. ***Interactive Derivation Tasks*** could include mini-games or exercises where learners build words from roots using affixes, reinforcing morphological rules and productive word formation strategies.

5. ***Gender- and Age-Aware Lexical Profiles*** could allow the tool to adapt to the learner's age and gender by offering customized lexical pathways. Insights from the corpus can be used to balance exposure to both expressive and functional vocabulary and offer developmentally appropriate lexical trajectories.

By embedding derivational insights into future lexicographic interfaces, we propose a framework for tools that not only reflect how language is used by learners but also actively shape how it is acquired, making lexicography not just descriptive, but pedagogically dynamic and developmentally transformative. With such tools we can foster morphological awareness as a foundation for vocabulary growth, support creative and stylistic development in writing, encourage equitable lexical exposure across learner groups, and bridge the gap between lexicographic resources and classroom practice.

7. Conclusion and future work

This study explored how derivational families rooted in CRT, PIS, and RAD are used in authentic learner writing, drawing on the ŠKOLARAC corpus of Croatian school essays. By combining corpus-based methods, morphological annotation in NooJ, and interactive visualizations in Tableau and Voyant Tools, we demonstrated how intelligent lexicographic tools can uncover subtle patterns in learner language, patterns that are often overlooked in traditional dictionary design.

Additionally, this study exemplifies the potential of intelligent lexicography as a methodological framework that goes beyond traditional corpus annotation and dictionary compilation. By integrating rule-based morphological parsing, root-based tagging, and interactive visualization tools, our approach allows for a dynamic, multidimensional analysis of learner language. This methodology not only facilitates the identification of derivational patterns but also supports a more learner-centered perspective, uncovering nuanced gender-based differences and usage trends that often remain invisible in conventional analyses. Such tools pave the way for developing adaptive, data-driven pedagogical resources and digital language applications tailored to authentic learner needs.

Our findings reveal that girls tend to use a broader range of derivational forms, while boys often rely on a smaller set of high-frequency lemmas. These differences are not merely quantitative but reflect distinct stylistic and thematic preferences. The integration of proportional lemma coverage, normalized frequency, TTR, and log-likelihood analysis provided a nuanced view of how derivational morphology functions in real-world learner production.

From a lexicographic perspective, this work demonstrates how intelligent lexicography can bridge the gap between empirical learner data and practical language resources. The tagging of root attributes, the extraction of derivational networks, and the visualization of gender-based variation all point toward a more dynamic, adaptive approach to lexical representation, one that reflects how language is actually used, not just how it is normatively defined.

Several directions emerge for future development. While this study focused on three roots, the methodology is scalable. Future work will include a broader set of derivational families to map the full morphological landscape of learner language. Blends and compounds like *radoholičar* 'workoholic' or *romanopisac* 'novelist', which combine multiple roots, were excluded from this analysis but represent fertile ground for exploring semantic compounding and lexical creativity. Also, the retention of misspelled forms opens the door to intelligent tools that recognize and respond to learner errors, not just by correcting them, but by understanding their morphological reasoning (morphological strategies), revealing how learners internalize and experiment with word-formation rules.

The derivational grammars and root-based tagging developed here can be embedded into digital dictionaries and writing assistants, offering real-time, context-sensitive lexical suggestions tailored to learner profiles. Applying the same methodology to learner corpora in other morphologically rich languages could reveal universal and language-specific patterns in derivational development. Ultimately, this work advocates for a lexicography that is not only intelligent in its architecture, but also empathetic in its orientation: attuned to the learner, responsive to variation, and grounded in authentic language use.

8. Acknowledgements

The research presented in this paper was conducted with partial financial support from the University of Zagreb (Short Term Grant No. 11-937-1036), contributing to the implementation of core linguistic and computational procedures.

9. References

Babić, S. (2002). *Tvorba riječi u hrvatskome književnome jeziku*. Zagreb: HAZU, Globus.

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V. & Znika, M. (2003). *Hrvatska gramatika*. Zagreb: Školska knjiga.

Filko, M., Šojat, K. & Štefanec, V. (2020). The design of CroDeriv 2.0. *The Prague Bulletin of Mathematical Linguistics*, 115, pp. 83–104.

Frankenberg-Garcia, A. (2014). How language learners can benefit from corpora, or not. In *Recherches en didactique des langues et des cultures [En ligne]*, 11-1. https://doi.org/10.4000/rdlc.1702

Gabrielatos, C. (2005). "Corpora and language teaching : just a fling or wedding bells?". *Teaching English as a Second Language Electronic Journal*, vol. 8, n° 4. pp. 1-35. Available at http://tesl-ej.org/ej32/a1.html

Hržica, G. (2021). Derivational morphology in Croatian child language. In *The Acquisition of Derivational Morphology. A Cross-linguistic Perspective*. Amsterdam: John Benjamins Publishing, pp. 141–168. https://doi.org/10.1075/lald

Kocijan, K. (2015). Visualizing natural language resources. In Pehar, F., Schlögl, C. & Wolff, C. (eds.) *Proceedings of the 14th International Symposium on Information Science* (ISI 2015). Zagreb: Verlag Werner Hülsbusch, pp. 203–216. Available at: https://zenodo.org/record/17934/files/s3_203-216.pdf

Kocijan, K. (2022). How we color the world with words. *Suvremena lingvistika*, 48(93), pp. 41–83. https://doi.org/10.22210/suvlin.2022.093.03

Kocijan, K., Janjić, M. & Librenjak, S. (2016). Recognizing diminutive and augmentative Croatian nouns. In *Automatic Processing of Natural-Language Electronic Texts with NooJ*. Cham: Springer, pp. 23–36.

Kocijan, K. & Šojat, K. (2024). Exposing diminutive and pejorative verbs in Croatian.

In Bartulović, A., Mijić, L. & Silberztein, M. (eds.) *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*. Cham: Springer, pp. 39–51. https://doi.org/10.1007/978-3-031-89810-5_4

Kocijan, K., Šojat, K. & Poljak, D. (2018). Designing a Croatian aspectual derivatives dictionary: preliminary stages. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing* (LR4NLP-2018). Stroudsburg (PA): Association for Computational Linguistics (ACL), pp. 28–37.

Kuna, Z. (2022). Tvorba imenica u ranom jezičnom razvoju na temelju podataka za jedno dijete iz Hrvatskog korpusa dječjeg jezika. *Suvremena lingvistika*, 48. doi: https://doi.org/10.22210/suvlin.2022.093.04

Lew, R. (2024). Dictionaries and lexicography in the AI era. Humanities and Social Sciences Communications. 11. https://doi.org/10.1057/s41599-024-02889-7.

Llach, M.P.A. (2010). Exploring the role of gender in lexical creations. In Catalán, R.M.J. (ed.) *Gender Perspectives on Vocabulary in Foreign and Second Languages*. London: Palgrave Macmillan. https://doi.org/10.1057/9780230274938_4

Murakami, A. & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3), pp. 365–401. https://doi.org/10.1017/S0272263115000352

Quixal, M., Rudzewitz, B., Bear, E. & Meurers, D. (2021). Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning* (NLP4CALL 2021). Linköping Electronic Conference Proceedings, 177, pp. 15–27.

Römer-Barron, U. (2023). Usage-based approaches to second language acquisition vis-à-vis data-driven learning. *TESOL Quarterly*, 58. https://doi.org/10.1002/tesq.3278

Silić, J. & Pranjković, I. (2005). *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.

Šojat, K., Kocijan, K. & Filko, M. (2019). Processing Croatian aspectual derivatives. In Mirto, M. et al. (eds.) *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*. Heidelberg: Springer, pp. 50–61.

Šojat, K. & Filko, M. (2023). Processing Croatian morphology: Roots, segmentation and derivational families. In Filko, M. & Šojat, K. (eds.) *Proceedings of the 4th International Workshop on Resources and Tools for Derivational Morphology*. Zagreb: HDJT, pp. 61–70.

Šojat, K., Srebačić, M. & Tadić, M. (2012). Derivational and semantic relations of Croatian verbs. *Journal of Language Modelling*, 0(1), pp. 111–142.

Šojat, K., Srebačić, M., Tadić, M. & Pavelić, T. (2014). CroDeriV: A new resource for processing Croatian morphology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14). Reykjavik: ELRA, pp. 3366–3370.

Tadić, M. & Fulgosi, S. (2003). Building the Croatian morphological lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*. Budapest: ACL, pp. 41–46.

Teibowei, M.T. (2024). Sociolinguistic variations and gender differences in language usage. *International Journal of English Language and Communication Studies*, 9(1), pp. 95–101.

Vučković, K., Librenjak, S. & Dovedan, Z. (2011). Deriving nouns from numerals. In *Proceedings of the NooJ 2010 International Conference and Workshop*. Komotini, pp. 84–95.

Vučković, K., Librenjak, S. & Dovedan Han, Z. (2013). Derivation of adjectives from proper names. In *Formalising Natural Languages with NooJ*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 57–68.

Vučković, K., Tadić, M. & Bekavac, B. (2010). Croatian language resources for NooJ. *CIT. Journal of Computing and Information Technology*, 18, pp. 295–301. https://doi.org/10.2498/cit.1001914