Title: Enhancing Croatian idiom recognition through contextual analysis

Applicant: Nika Marinović

Affiliation: University of Zagreb, Faculty of humanities and social sciences, Croatia

Keywords: Croatian language, idioms, contextual analysis, lexical approach

Description:

The proposed project aims to develop a comprehensive resource for enhancing the recognition of Croatian idioms through a combination of lexical and contextual analysis. Building upon the traditional lexical approach, this project will integrate modern contextual analysis techniques to improve the accuracy and effectiveness of idiom identification in Croatian texts. The project will begin by compiling a curated list of Croatian idioms and their variations, focusing on capturing their meanings and usage patterns along with their contextual usage in authentic texts. By leveraging contextual meaning in the analysis of idioms, the project seeks to provide valuable insights into the linguistic characteristics, cultural significance, and nuanced understanding of idiom usage in different linguistic contexts within Croatian language usage.

The project will commence with an extensive compilation of a diverse range of Croatian idioms and their variations. These idiomatic expressions will be sourced from authentic texts, literature, and spoken language, ensuring a comprehensive representation of idiomatic usage in various contexts. For instance, idioms like "biti u sedmom nebu" (to be on cloud nine) or "ljutiti se kao ris" (to be very angry) are commonly used in Croatian, each carrying its unique cultural connotations. Following this analysis, a user-friendly database or online platform will be developed, where users can access these idiomatic expressions alongside their meanings, usage examples, and cultural insights. This platform serves as a valuable resource for language learners, translators, and anyone interested in gaining a deeper understanding of Croatian language and culture through its idiomatic expressions.

In summary, the project aims to not only compile a comprehensive collection of Croatian idioms but also to make them accessible and understandable to a wider audience. Through documenting and analyzing these idiomatic expressions, the project seeks to promote the preservation and appreciation of Croatian language and culture.

Participating in the UniDive Training Summer School 2024 holds significant importance for the advancement of this project. Through engagement in specialized training sessions focusing

on language technology and corpus annotation techniques, the project team will acquire invaluable practical knowledge and essential skills crucial for the successful development of the Croatian idiom recognition resource. These sessions will offer insights into the latest advancements and best practices in linguistic annotation, enabling the team to effectively annotate idiomatic expressions with lexical and contextual information.

Furthermore, the brainstorming hackathon sessions present a unique opportunity for the project team to collaborate with peers and experts in the field. These sessions will serve as a platform for refining the project's methodology, discussing innovative approaches, and addressing any open questions or challenges encountered during the development process. Through collaborative problem-solving and knowledge-sharing, the project team will be able to enhance the accuracy and effectiveness of the idiom recognition system, ensuring its alignment with the evolving needs and complexities of Croatian language usage.

Open questions for brainstorming hackathon:

1. How can machine learning techniques be integrated into the idiom recognition process to improve accuracy?

2. What strategies can be employed to handle idiomatic expressions with regional variations or dialectal differences?

3. How can the idiom recognition system adapt to evolving language usage and new idiomatic expressions over time?

Project phase: Planning

This approach combines the traditional lexical approach with contextual search to enhance the recognition of Croatian language idioms. Integrating contextual meaning into idiom analysis can help improve the precise identification of idioms and their variations in real texts. Using curated lists of idioms and their variations as a basis for idiom identification provides a solid foundation for analysis, while contextual search offers deeper understanding of idiom usage in different situations. This integrated approach holds the potential to develop a more effective idiom recognition system that accurately reflects their usage in the Croatian language.

References:

1) Kovačević, B. i Ramadanović, E. (2013). Frazemske polusloženice (od rječnika preko tvorbe do pravopisa i obratno). *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 39* (1), 271-291. Retrived from https://hrcak.srce.hr/113037
2) Omazić, M. (2002). O poredbenom frazemu u engleskom i hrvatskom jeziku. *Jezikoslovlje, 3* (1-2), 99-129. Retrived from https://hrcak.srce.hr/31348
3) Filipović Petrović, I. i Parizoska, J. (2019). Konceptualna organizacija frazeoloških rječnika u e-leksikografiji. *Filologija,* (73), 27-45. Retrived from https://hrcak.srce.hr/230913
4) Gašparović, M. (2023). Translating English-Croatian Idioms: Where Language and technology Meet. Brala-Vukanović, M. (mentor). Rijeka, Sveučilište u Rijeci. Retrieved from https://croris.hr/crosbi/publikacija/ocjenski-rad/797957
5) Tadić, M. (2001). Building the Croatian Language Technologies Portal.
6) Senaldi, M., Titone, D., Johns, B. (2022). Determining the Importance of Frequency and Contextual Diversity in the Lexical Organization of Multiword Expressions. Canadian Journal of Experimental Psychology. 76. Retrieved from https://www.researchgate.net/publication/358524589_Determining_the_Importance_of_Frequency_and_Contextual_Diversity_in_the_Lexical_Organization_of_Multiword_Expressions
7) Dunđer, I. (2020). Machine Translation System for the Industry Domain and Croatian Language. *Journal of Information and Organizational Sciences, 44* (1), 33-50. Retrived from https://doi.org/10.31341/jios.44.1.2
8) Kocijan, K., Požega, M. i Poljak, D. (2016). Scholarly reference trees. *Libellarium, 9* (2), 0-0. Retrived from https://hrcak.srce.hr/176678
9) Slamić Tarade, S. (2020). Istraživanje branda korištenjem analize teksta i konteksta. *Polytechnic and design, 8* (2), 74-82. Retrived from https://doi.org/10.19279/TVZ.PD.2020-8-2-02
10) Narasimhan, R. (2021). Reading idioms with natural language processing. LinkedIn. Retrieved from https://www.linkedin.com/pulse/reading-idioms-natural-language-processing-ram-narasimhan
11) Salton, G. (2017) Representations of Idioms for Natural Language Processing: Idiom type and token identification, Language Modelling and Neural Machine Translation. Doctotal thesis, DIT. Retrived from https://doi.org/10.21427/D77H8K