The RomCro parallel corpus v.2.0. and its application in the contrastive analysis of infinitival and finite complement constructions

Key words (5): parallel corpus; Romance languages; Croatian; collostructional analysis; complement constructions

Abstract:
This presentation introduces RomCro v.2.0, the latest version of a multilingual, multidirectional parallel corpus which includes contemporary literary texts in six Romance languages and Croatian and is available on platforms Sketch Engine and HR-CLARIN. Building on the foundation of RomCro v.1.0 (Authors, 2023a), which included Spanish, French, Italian, Portuguese, Romanian, and Croatian, the updated version expands the corpus with additional texts and a new language, Catalan. This expansion has increased the corpus to 19.4 million words, further enhancing its value as a resource for philological and linguistic research—including contrastive, lexicographic, phraseological, glottodidactic, and translation studies (e.g., Granger, Lerot, & Petch-Tyson, 2003; Teubert, 2007)—as well as for machine translation training (Koehn, 2020), translator education (López Rodríguez, 2016), and terminology extraction (Lefever, Macken, & Hoste, 2009). RomCro has been presented on various occasions and has proven to be a valuable resource for different types of contrastive analyses (e.g. on the use of articles in Romance languages (Authors, 2023b) or noun determination (Author, 2020)) and students have successfully applied it in seminar and thesis work.

In this presentation, we focus on one specific application of RomCro: the contrastive analysis of constructions with infinitival and finite complements in Romance languages and Croatian. More specifically, this study employs Cognitive Construction Grammar (Boas, 2013) and corpus linguistics methodology to examine postverbal, non-prepositional infinitive and finite complements. Due to space limitations and structural differences among the languages, our analysis primarily focuses on Spanish, French, and Croatian, considering only verbs that can potentially alternate between both constructions (as illustrated in (1), (2) or (3), depending on the language). Verb-complement combinations are treated as schematized constructions and are analyzed using distinctive collexeme analysis, a method within collostructional analysis that identifies verbs most strongly associated with either type of complementation (Stefanowitsch & Gries, 2003; Gries & Stefanowitsch, 2004). Additionally, the results of the collostructional analysis are qualitatively examined from both morphosyntactic and semantic perspectives, taking into account key factors such as verbal mood (indicative vs. subjunctive) and verbal aspect (in Croatian), coreferentiality between the subjects of the main and complement verbs, and verb class membership—all of which have been shown to play a crucial role in these alternations (e.g., Yoon & Wulff, 2016; Kaleta, 2023). Based on our preliminary research, we expect to find that verbs allowing both infinitival and finite complements exhibit a probabilistic distribution, influenced by the aforementioned factors, rather than forming mutually exclusive categories. Furthermore, our findings reveal broader tendencies in Romance languages on the one hand and highlight specific syntactic and semantic choices in Croatian as a Slavic language on the other, thus contributing to a more thorough description of French and Spanish, as well as Croatian.

Examples:

(1) A. Sp. *Creía saberlo todo sobre aquella mujer* [...].
B. Fr. *Je croyais tout connaître de cette femme* [...].
C. Cr. *Mislio sam da o toj ženi sve znam* [...].

(2) A. Sp. [...] *creí que sabía quién era* [...].
B. Fr. [...] *j'ai cru savoir qui il était* [...].
C. Cr. [...] *mislio sam da znam tko je on* [...].

(3) A. Sp. [...] *no creo que supieran que estaban fusilando a uno de los fundadores de Falange* [...].
B. Fr. [...] *je ne crois pas qu'ils aient su qu'ils étaient en train d'exécuter l'un des fondateurs de la Phalange* [...].
C. Cr. [...] *sumnjam da su znali da strijeljaju jednog od osnivača Falange* [...]

References:

1. Author Citation (2020).
2. Authors Citation (2023a).
3. Authors Citation (2023b).
4. Boas, H.C. (2013). Cognitive Construction Grammar. In Hoffmann, Th. & Trousdale, G. (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 233-252). Oxford University Press.
5. Granger, S., J. Lerot, & Petch-Tyson, S. (Eds.). (2003). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Rodopi.
6. Gries, S. Th. & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on ‘alternations'. *International journal of corpus linguistics, 9*(1), 97-129.
7. Kaleta, A. (2023). The semantics of clausal complementation: Evidence from Polish. *Journal of Slavic Linguistics, 31*(1), 99-132.
8. Koehn, P. (2020). *Neural Machine Translation*. Cambridge University Press.
9. Lefever, E., L. Macken, & Hoste, V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL* (pp. 496-504). Association for Computational Linguistics.
10. López Rodríguez, C. I. (2016). Using Corpora in Scientific and Technical Translation Training: Resources to Identify Conventionality and Promote Creativity. *Cadernos de tradução, 1*, 88-120.
11. Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics, 8*(2), 209-243.
12. Teubert, W. (Ed.). (2007). *Text Corpora and Multilingual Lexicography*. John Benjamins Publishing Company.
13. Yoon, J. & Wulff, S. (2016). A corpus-based study of infinitival and sentential complement constructions in Spanish. In Yoon, J. & S. Th. Gries (Eds.), *Corpus-Based Approaches to Construction Grammar* (pp. 145-164). John Benjamins.