Mind the Gap: Contrastive Computational Assessment of Crowd-Sourced Linguistic Knowledge on Morphological Gaps in Latin and Italian

Keywords: Morphology; Defectivity; Wikipedia; Latin; Italian

When asked what the past tense for *forgo* is, GPT-4o responds, "The past tense of 'forgo' is forwent." Yet, most native speakers would find *forwent* unacceptable but are unable to find the right, natural form for the past tense of *forgo* [Gorman and Yang, 2019]. Words such as *forwent* are morphological gaps in which expected forms are absent, an intriguing and underexplored linguistic phenomenon that "remains poorly understood" [Baerman and Corbett, 2010]. Documenting defectivity typically requires significant expertise and manual effort, making crowd-sourced content a potentially valuable but underexplored resource. Wikipedia and its sister websites, prominent examples of user-generated content, rank consistently among the most popular websites worldwide, attracting over 4.5 billion monthly visitors. Despite their extensive reach and usage, they are generally perceived as unreliable by domain experts. However, for scarce linguistic phenomena such as defectivity in non-English languages, Wikipedia and Wiktionary often serve as two of the few widely accessible and frequently utilized resources.

In this study, we conduct computational analyses of inflectional gaps by using UDPipe [Straka et al., 2016] to tokenize raw text and customizing UDTube [Yakubov, 2024], a state-of-the-art neural morphological analyzer trained with Universal Dependencies (a corpus of morphologically annotated text in different languages), to incorporate mBERT [Conneau et al., 2020] as an encoder and annotate large corpora of text in Latin (640MB, 390 million words) and Italian (8.3GB, 5 billion words). The resulting massive annotated data are then used to measure the frequency of inflectional forms and validate lists of defective verbs scraped and compiled from Wiktionary's Latin and Italian pages to verify which verbs are confirmed computationally to be morphological gaps. The aim is to conduct quality assurance on crowd-sourced linguistic information provided in Wiktionary, assessing and comparing its reliability for linguistic knowledge in two related, yet distinct languages and flagging cases where lemmata are labeled as defective but may not actually be defective, based on linguistic evidence.

We employ two models to quantify the likelihood of non-defectivity. The first is **absolute frequency**. The second is **divergence from expected frequency**–how far a given inflected word diverges from its expected probability. To measure this, we use the **log-odds ratio** [Gorman and Yakubov, 2024].

Our findings indicate that nearly 70% of Latin and 80% of Italian gaps listed in Wiktionary strongly align with computational results, thus suggesting a degree of reliability in Wiktionary's linguistic data, despite coming from unreferenced, user-generated sources. However, 7% of Latin lemmata labeled as defective show strong corpus evidence of being non-defective, compared to less than 4% in Italian. This finding raises some concerns about potential limitations of crowd-sourced wikis as definite sources of linguistic knowledge, despite their value as resources for rare linguistic features. This work contributes to contrastive morphological research by computationally comparing morphological defectiveness in Latin and Italian. We provide a novel, scalable methodology for validating and expanding knowledge of morphological defectivity in these two historically related yet distinct languages. Through this contrastive computational approach, we contribute to advancement of knowledge and discussion on morphology, defectiveness, and the role of digital resources in linguistic research.

References

Matthew Baerman and Greville G. Corbett. 2010. *Defective Paradigms: Missing Forms and What They Tell Us*. Oxford University Press, Oxford.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kyle Gorman and Daniel Yakubov. 2024. Acquiring inflectional gaps with indirect negative evidence: evidence from russian. In *Proceedings of the 55th Annual Meeting of the North East Linguistic Society (NELS 55)*.

Kyle Gorman and Charles Yang. 2019. When nobody wins. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors, *Competition in Inflection and Word-Formation*, pages 169–193. Springer Cham.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Daniel Yakubov. 2024. *How Do We Learn What We Cannot Say?* Ph.d. dissertation, City University of New York (CUNY). Available at CUNY Academic Works.