Mind the Gap: Computational Quality Assurance of Crowd-Sourced Linguistic Knowledge on Latin and Italian Morphological Gaps

Jonathan Sakunkoo and **Annabella Sakunkoo**
Stanford University OHS
{jonkoo,apianist}@ohs.stanford.edu

1 Introduction

> The past tense of "forgo" is forwent. So, you would say: "I forwent this position." It's a bit formal or uncommon in modern usage, but grammatically correct.

Above is a response from GPT-4o when asked what the past tense for "forgo" is. Yet, most fluent English speakers would find *forwent* unnatural, ineffable (Gorman, 2023), and unacceptable (Embick and Marantz, 2008). Most English speakers would also be unable to find the right, natural form for the past tense of *forgo* (Gorman and Yang, 2019). Words such as *forgo* are instances of defective verbs or morphological gaps in which expected forms are absent—a problematic intrusion of morphological idiosyncrasy (Baerman and Corbett, 2010).

While inflectional gaps are not a recently discovered phenomenon, they "remain poorly understood" (Baerman and Corbett, 2010) and documenting them requires extensive human expertise and effort. For scarce linguistic phenomena in less-studied languages, Wikipedia and Wiktionary serve as among the few widely accessible and frequently utilized resources, consistently ranked among the most popular websites globally. With its extensive reach and usage, crowd-sourced content is a potentially valuable but underexplored resource although its user-contributed nature has sparked controversy on its overall trustworthiness.

In this study, we conduct computational analyses of inflectional gaps by customizing UDTube (Yakubov et al., 2024), a scalable state-of-the-art neural morphological analyzer trained with Universal Dependencies (a collection of corpora of morphologically annotated text in different languages), to incorporate mBERT (Devlin et al., 2019) as an encoder and annotate large corpora of text in Latin and Italian (Conneau et al., 2020). The resulting massive annotated data are then used to measure the frequency of certain inflectional forms of interest 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 inflectional gaps.

By bridging computational techniques with linguistic analysis, the study contributes to linguistics of less-explored languages and offers novel insights and computational methodologies for scalable quality assurance and validation of crowd-sourced content, while addressing gaps in linguistic knowledge.

2 Data

This study uses Universal Dependencies (UD), Common Crawl, and Wiktionary in the computational validation of morphological gaps. Universal Dependencies is a collection of multilingual treebanks for syntactic and morphological analysis across languages (Nivre et al., 2017). We utilize the largest available treebanks for Italian and Latin in the UD dataset. For corpora, we use an 8.3GB dataset containing approximately 5 billion tokens of diverse Italian text and a 640MB dataset with approximately 390 million tokens of Latin text.

3 Methods

As shown in Figure 1, this study uses a computational approach to validate inflectional gaps in Latin and Italian in three major steps: (1) Training UDTube with Universal Dependencies, (2) Annotating Large-Scale Text with UDTube[1], and (3) Validating Defective Forms.

4 Results and Conclusion

In the evaluation of defective lemmata listed in Wiktionary against corpus evidence, lemmata are classified into likely defective (based on expert-recommended frequency threshold of 10), on the edge, and likely not defective.

[1]The tuned morphological analyzer achieves 98% and 96% accuracy on the Latin and Italian test sets, respectively.

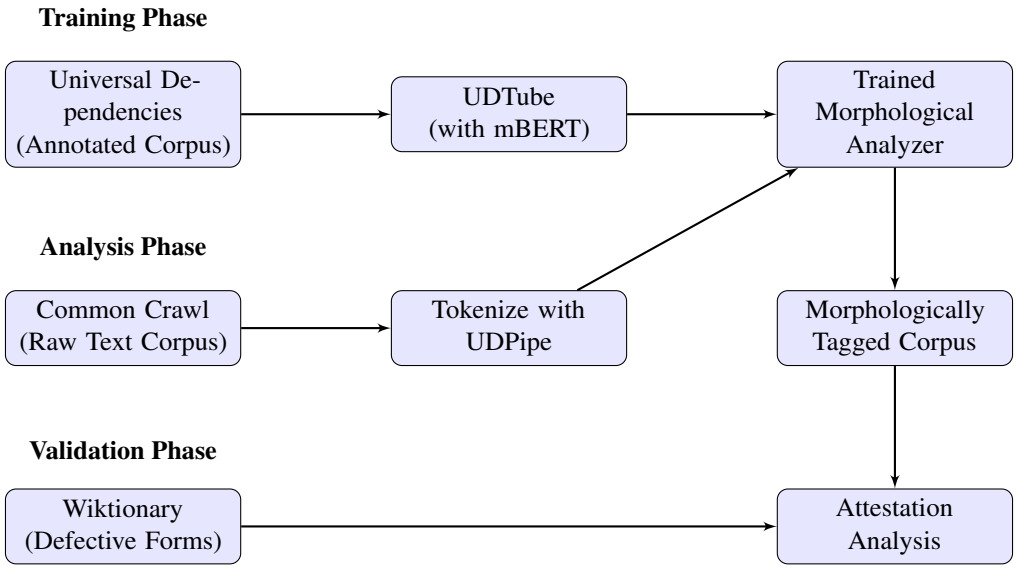

Figure 1: Workflow for computational validation of morphological gaps, using UDTube

Occurrences	Latin	Italian
Likely defective: ≤ 10	67.4%	79%
On the edge: $11 - 100$	25.4%	17%
Likely not defective: > 100	7.2%	4%

Table 1: Summary of defective forms in Wiktionary

Based on this result, Wiktionary's list of defective verbs in Italian is 1.8 times less likely to contain errors compared to Latin. The computational results, together with manual verification by human experts, suggest that while Wiktionary provides a reliable account of Italian morphological gaps, at least 7% of Latin lemmata listed as defective are unlikely to be truly defective. This discrepancy highlights potential limitations of crowd-sourced wikis as definitive sources of linguistic knowledge, particularly for less-studied phenomena and languages, despite their value as resources for rare linguistic features. This study presents a novel computational approach to validating defectivity in a crowd-sourced linguistic resource and contributes to expanding our morphological knowledge.

5 Acknowledgement

We are grateful to Kyle Gorman for valuable advice and to Yale NENLP researchers, Dan Jurafsky, and reviewers for insightful feedback for future work.

References

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

Jeremy K. Boyd and Adele E. Goldberg. 2011. Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, 87(1):55–83.

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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Embick and Alec Marantz. 2008. Architecture and blocking. *Linguistic Inquiry*, 39(1):1–53.

Kyle Gorman. 2023. *Morphological Defectivity*.

Kyle Gorman and Charles Yang. 2019. *When Nobody Wins*. Springer International Publishing.

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.

Daniel Yakubov, Kyle Gorman, and Github Contributor Jonathan Sakunkoo. 2024. UDTube: A tool for universal dependency-based linguistic analysis.