Lost and Found: Computational Quality Assurance of Crowdsourced Knowledge on Morphological Defectivity in Wiktionary

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

Abstract

Morphological defectivity is an intriguing and understudied phenomenon in linguistics. Addressing defectivity, where expected inflectional forms are absent, is essential for improving the accuracy of NLP tools in morphologically rich languages. However, traditional linguistic resources often lack coverage of morphological gaps as such knowledge requires significant human expertise and effort to document and verify. For scarce linguistic phenomena in under-explored languages, Wikipedia and Wiktionary often serve as among the few accessible resources. Despite their extensive reach, their reliability has been a subject of controversy. This study customizes a novel neural morphological analyzer to annotate Latin and Italian corpora. Using the massive annotated data, crowd-sourced lists of defective verbs compiled from Wiktionary are validated computationally. Our results indicate that while Wiktionary provides a highly reliable account of Italian morphological gaps, 7% of Latin lemmata listed as defective show strong corpus evidence of being non-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. By providing scalable tools and methods for quality assurance of crowd-sourced data, this work advances computational morphology and expands linguistic knowledge of defectivity in non-English, morphologically rich languages.

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. Similarly, Llama 3.2 confidently replies that

> The past tense of "forgo" is "forwent".

Yet, most English speakers would find *forwent* ineffable (Gorman, 2023) and unacceptable (Embick and Marantz, 2008). Most English speakers are actually unable to find the right, natural form for the past tense of *forgo* (Gorman and Yang, 2019). Similarly, *beware* functions exclusively as a positive imperative (e.g. beware the bear!), and *BEGO* can only appear as the imperative *begone!* Words such as these are instances of defective verbs or morphological gaps in which expected forms are missing—a problematic intrusion of morphological idiosyncrasy (Baerman and Corbett, 2010). In other words, a lexeme is defective if at least one of its possible inflectional variants is ineffable (Gorman, 2023) or exhibits relative non-use (Sims, 2006).

In Latin, *aiō* 'to speak' lacks the first- and second-person plural present forms. Another defective verb is *inquam* 'to say', also restricted to an incomplete subset of forms, such as the third person singular in the present and perfect indicative (e.g. *inquit*) (Oniga and Shifano, 2014).

While inflectional gaps are not a recent discovery, they "remain poorly understood" (Baerman and Corbett, 2010). Since NLP systems often assume regular paradigms, accounting for defectivity would improve the accuracy so as to not use or suggest forms that do not exist, especially for less-studied and morphologically rich languages where inflectional gaps are more common. Gorman and Yakubov (2024) applied UDTube to discriminate defective from non-defective words in Russian and Greek. While curated lists of defective verbs exist for languages such as Russian and Greek, verified resources remain scarce for many others, including Latin and Italian. For scarce linguistic phenomena in less-studied languages, Wikipedia and Wiktionary often serve as widely accessible and frequently utilized resources, consistently ranked among the most popular websites globally, attract-

ing over 4.5 billion monthly visitors. With extensive reach and usage, crowd-sourced content is a potentially valuable resource; projects like UniMorph (Kirov et al., 2018) have extracted morphological data from Wiktionary. However, despite its many virtues, its crowdsourced nature has sparked controversy on trustworthiness and reliability.

In this study, we conduct computational analyses of inflectional gaps by customizing UDTube (Yakubov, 2024)[1], 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. We apply this enhanced model to 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 validate lists of defective verbs scraped and compiled from Wiktionary's Latin and Italian pages to verify which verbs are confirmed computationally to be defective or non-defective.

We model defectivity after how children might learn what the gaps or defective forms are–in other words, learn what is missing. Brown and Hanlon (1970) showed that parents typically provide explicit feedback on the truth value of a child's articulation but rarely correct grammatical errors, such as inflection, thus implying that children do not acquire morphology through explicit negative evidence. Similarly, Baronian (2005) reinforced the idea that morphological gaps are not taught directly. While the exact process by which children acquire defectivity remains unclear, many scholars in linguistics and language learning agree that gaps are primarily learned through **Indirect** (or implicit) **Negative Evidence** (INE) (Orgun and Sprouse, 1999; Johansson, 1999; Sims, 2006).

Our findings indicate that nearly 80% of inflectional gaps in Italian and 70% in Latin listed in Wiktionary strongly align with our computational INE results while 4% of Italian and 7% of Latin lemmata labeled as defective in Wiktionary show a high tendency to actually be non-defective, thus suggesting a degree of reliability in Wiktionary's linguistic data, despite coming from unreferenced, user-generated sources. The study also identifies multiple inaccuracies, particularly in Latin, and highlights the need for more rigorous expert verification in crowd-sourced linguistic resources.

[1] https://github.com/CUNY-CL/udtube

This study explores the potential and limitations of crowd-sourced content as a supplementary linguistic resource. By using a novel, scalable approach for computationally analyzing morphological gaps, it advances the intersection of computational methods and linguistics as it contributes to quality assurance of crowdsourced content and addresses gaps in linguistic knowledge.

2 Data

We employ the following data sources in the computational validation of morphological gaps.

Universal Dependencies (UD) (Nivre et al., 2017): We utilize two of the largest available Latin and Italian treebanks—UD Latin ITTB and UD Italian VIT—to train our morphological analyzer.

Common Crawl (CC-100) (Wenzek et al., 2020): From CC-100, we use an 8.3GB dataset containing 5B tokens of Italian text and a 640MB dataset with over 390M tokens of Latin text.

Wiktionary: We scrape and compile lists of defective verbs and inflectional gaps from Latin and Italian pages of Wiktionary. This study focuses on Latin and Italian because of their reasonably large number of inflectional gaps and their representation in Wiktionary, which contains the most extensive lists of morphological gaps for these languages.

3 Methodology

As shown in Figure 1, this study uses a computational approach to validate inflectional gaps in Latin and Italian in three major steps:

Training UDTube with UD: As a neural morphological analyzer, UDTube's primary purpose is to decompose words morphologically and identify their morphological features. We trained UDTube using the mBERT encoder, a multilingual BERT model trained on 104 languages (Devlin et al., 2019), on the UD Italian and Latin treebanks. UDTube has been demonstrated to have superior performance in recent comparative studies (Yakubov, 2024), which show that it achieves high accuracy in morphological annotations, outperforming the popular UDPipe (Straka et al., 2016) in multiple languages. Our tuned UDTube model has 98% and 96% accuracies in Features Morphological Annotations in Latin and Italian, respectively.

In hyperparameter tuning, optimal hyperparameters were determined using Weights and Biases, a tool for tracking and visualizing experiments. This

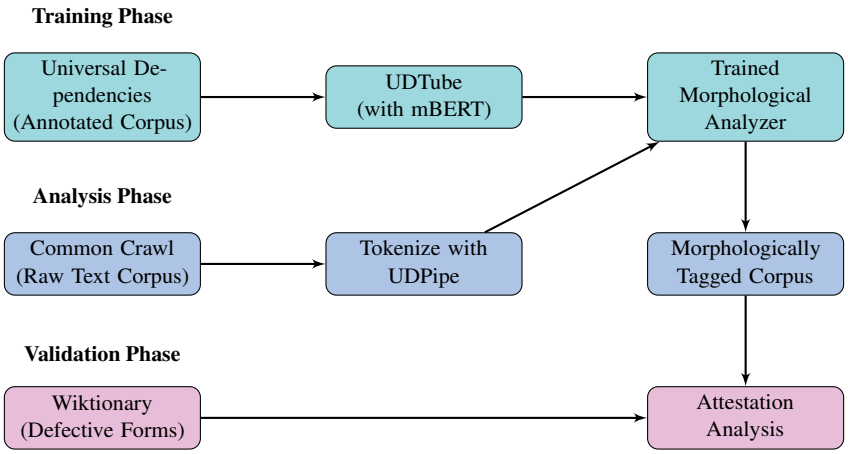

Figure 1: Workflow for computational validation of morphological gaps, using UDTube

step ensured that UDTube's configuration was fine-tuned for Latin and Italian datasets.

Annotating Large-Scale Text: The trained UDTube model is used to annotate text from the Common Crawl corpora. The process involved:

- **Text Preprocessing:** The raw text was cleaned and tokenized using UDPipe (Straka et al., 2016) into words.
- **Morphological Tagging:** Each token was analyzed and annotated with its lemma and morphological features, using the trained UDTube model. This produced a morphologically tagged corpus in CoNLL-U format.
- **Frequency Database:** From the tagged data, we generated a frequency database containing the occurrence counts for each morphological form of every lemma.

Validating Defective Forms: To verify the defective forms listed in Wiktionary, we applied the principle of **Indirect Negative Evidence** (Gorman and Yang, 2019; Boyd and Goldberg, 2011), a key mechanism in language acquisition by which learners infer defectivity: if a certain morphological form is defective, then it should not occur or occur extremely infrequently in usage. We employ two models to quantify the likelihood of non-defectivity. The first is **absolute frequency**. If a possible word has a high absolute frequency, it is unlikely to be defective. The second is **divergence from expected frequency**. If the frequency of a possible inflected word is significantly higher than expected, assuming all else is equal, it is unlikely to be defective.

For each attested inflected word w, there exist a corresponding lemma l and a morphosyntactic feature bundle f. Let p_w, p_l, and p_f denote the probability of a word, lemma, and feature bundle, respectively, calculated from maximum likelihood estimation using corpus frequencies. Assuming independence and all else equal, p_w should be in proportion to $p_l \cdot p_f$. To measure **divergence from expected frequency**, how far a given inflected word has diverged from its expected probability, we use the **log-odds ratio** (Gorman and Yakubov, 2024).

$$\text{Log-odds ratio: } L_w = \log\left(\frac{p_w}{p_l \cdot p_f}\right)$$

The log-odds ratio has been found to be the best unexpectedness predictor for acceptability judgment. A log-odds ratio of 1.9 or more is considered to indicate a large divergence (Chen et al., 2010).

The reliability of Wiktionary's crowd-sourced data was assessed by calculating the percentage of purported defective forms that aligned with our computational findings. The evaluation was grouped into true positives, which are cases where the Wiktionary-listed defective form was confirmed as absent or extremely rare in the corpus, and false positives, which are cases where a supposedly defective form was frequently attested in the corpus, indicating an error in Wiktionary. For discrepancies, we conducted manual reviews to determine whether they arose from corpus limitations, UDTube errors, or inaccuracies in Wiktionary.

4 Results

In evaluating defective lemmata listed in Wiktionary against corpus evidence, lemmata are classified into four groups:

Not Attested: No inflected form of the lemma appears in the corpus, so we cannot confidently

verify whether it is defective or not. These lemmata are excluded from our analysis.

Likely Defective: The lemma's alleged defective form occurs ≤ 10 times in the corpus, indicating significant rarity, non-use, or absence.

On the Edge: The lemma's alleged defective form occurs 11-100 times in the corpus.

Attested but Not Defective: The lemma's forms occur frequently in the corpus, suggesting usage despite being listed as inflectional gaps in Wiktionary.

Occurrences	Latin	Italian
Likely defective: ≤ 10	67.4%	79.2%
On the edge: 11 - 100	25.4%	17.0%
Likely not defective: > 100	7.2%	3.8%

Table 1: Validation of Wiktionary's defective verbs

Log-Odds Ratio	Latin	Italian
> 1.9	6.3%	0.0%
> 1.5	12.2%	5.9%

Table 2: Verbs found to be likely non-defective due to very high p_w relative to $p_l \cdot p_f$

As shown in Table 1, Wiktionary's list of defective verbs in Latin is 1.8 times more likely to contain errors compared to Italian. This may be due to (1) the larger number of contemporary Italian speakers, leading to a stronger collective understanding of the language, and (2) Italian's less complex inflectional system compared to Latin. Table 2 shows the percentages of purported defective verbs that appear very frequently, relative to expected frequency. Based on the Log-Odds Ratio model and the threshold of large divergence (Chen et al., 2010; Cohen, 2013), approximately 6.3% of Latin lemmata labeled as defective in Wiktionary may actually be non-defective. Similarly, the absolute frequency measure indicates that approximately 7% of Wiktionary-listed defective Latin verbs are highly likely to be non-defective.

4.1 Discussion of Latin Results

For Latin, 1,190 defective lemmata are sourced from Wiktionary. Of these, 1,050 lemmata (88%) are attested in the corpus. Among the attested lemmata, 67% exhibit defective behavior (i.e., some forms suggested by Wiktionary are verified to have extremely low frequencies). For example, *discrepo* 'to disagree' is a defective lemma. Wiktionary claims that *discrepo* lacks a passive voice, and we found *discrepo* to occur only 3 times in the passive voice. However, *excommunico* 'to excommunicate' is an example of Attested but Not Defective Lemmata as it is claimed by Wiktionary to lack a perfect aspect but actually has a perfect form that occurs 846 times. Examples of Not Attested Lemmata are *astrifico*, *superfulgeo*, and *auroresco*.

4.2 Discussion of Italian Results

For Italian, 124 defective lemmata are obtained from Wiktionary, and 103 (83%) are attested in the corpus. Of the attested lemmata, 79% exhibit defective behavior. For example, *vèrtere* 'to concern' occurs 6 times in the past participle form, below the threshold of 10, corroborating Wiktionary's claim that *vèrtere* has no past participle form.

Our system identifies potential candidates for errors in Wiktionary, such as *consumere* 'to consume', *concernere* 'to concern', and *malandare* 'to be ruined'. For example, some native speakers confuse *consumere* with *consumare* 'to consume' (sometimes mistakenly perceiving the word as a more formal variant). Thus, although *consumere* is an archaic remnant from Latin and is listed on Wiktionary as defective and nonexistent in modern Italian, it is in fact still occasionally found to be in use. *Ludendo* 'playing' is another word detected by our model to be unlikely to be defective as *ludendo* appears frequently in the corpus due to code-switching with Latin.

5 Conclusion

This study presents a novel computational approach for quality assurance of a widely used crowd-sourced linguistic resource. Our findings highlight the potential and limitations of crowd-sourced linguistic references while demonstrating the effectiveness of scalable NLP models, such as UDTube, in verifying morphological gaps in less-studied languages. The results indicate that Wiktionary is a reasonably reliable resource, with limitations. This study hence illustrates the importance of computational validation for crowd-sourced linguistic data as the results show that some verbs marked as defective in Wiktionary are, in fact, functional and widely used. Moreover, the differences between Italian and Latin results suggest that linguistic evolution and corpus representativeness may impact the reliability of crowd-sourced morphologi-

cal knowledge. Latin exhibits more inconsistencies, thus highlighting the need for careful interpretation of crowd-sourced knowledge and corpus-based evidence in the absence of native speakers.

Future research can expand upon this work by extending the methodology to other languages to assess the completeness and accuracy of crowd-sourced resources. Beyond defective verbs, this approach can also be applied to other linguistic features, while integrating more diverse corpora, improving neural morphological analyzers, and experimenting with thresholds could enhance the ability to distinguish rare but valid forms from true gaps.

By bridging computational methods with linguistic inquiry, our novel empirical results demonstrate how NLP can enhance the quality assurance of crowdsourced linguistic resources. The study also uniquely contributes to expanding linguistic databases and our understanding of language structure across typologically diverse systems.

6 Limitations

Future work could explore whether models like XLM-RoBERTa provide more accurate results than mBERT for Latin and Italian. The corpora also have some limitations, particularly in Latin, as certain verb forms may be underrepresented or entirely absent. Since corpus coverage for Latin is inherently limited, some rare but valid inflectional forms may exist in texts outside the dataset. This incompleteness may contribute to false positives in our classification of defectivity, affecting the accuracy of frequency-based and statistical assessments. Additionally, context and pragmatics influence defectivity—some verbs classified as defective may still function within specific dialects, historical periods, or contexts. Furthermore, since no standardized thresholds exist for determining defectivity, our criteria remain somewhat arbitrary. These limitations suggest that while corpus analysis provides valuable insights into the functional status of defective verbs, it should be supplemented with qualitative linguistic expertise and historical context.

Another way that results may be impacted is the accuracy of UDTube. As expected from any models, UDTube is not perfect. Acknowledging that the annotation of morphological characteristics (FEATS) remains challenging, we chose UDTube due to its demonstrated superior performance in comparative studies (Yakubov, 2024). Our tuned UDTube model achieved 96% accuracy on the Italian holdout test set and 98% accuracy on the Latin holdout test set. Future work may further measure the performance of morphological analyzers in recent shared tasks, such as EvaLatin (Sprugnoli et al., 2022), to advance evaluation standards for morphological analysis. Additionally, as annotating the corpora is a computationally intensive task, we used distributed computing to complete the tagging in a reasonable timespan. Along the way, some nodes failed to complete their task, leaving some parts of the corpora untagged. Some cases of the limitations addressed above may have been avoided had the remaining portion of the corpora been used, but this is likely insignificant.

Finally, this study is descriptive rather than prescriptive. Our goal is not to prescribe what forms should or should not exist but to assess the degree to which a widely used crowd-sourced resource (e.g. Wiktionary) aligns with large-scale corpus evidence. Our computational models are designed for empirical evaluation, not to prescribe correctness. As such, our findings should be viewed as tools to support and refine linguistic understanding, particularly for under-documented phenomena. Similarly, when we refer to native speakers or expert verification, we do so not to invoke authority, but to acknowledge the limitations of corpus data and crowd-sourced data. We therefore view computational models, corpus data, crowd-sourced resources, and linguistic expertise as complementary: each contributes to a more robust and nuanced descriptive account of defectivity, especially in historically complex languages like Latin and Italian.

Acknowledgments

We are grateful to Kyle Gorman for his valuable guidance, advice, and support throughout this work. We also thank the Yale University NENLP'25 researchers, Dan Jurafsky, and the anonymous reviewers for their insightful feedback on future directions.

References

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

Luc V. Baronian. 2005. *North of phonology*. Phd dissertation, Stanford University, Department of Linguistics.

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.

Roger Brown and Camille Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In John R. Hayes, editor, *Cognition and the development of language*, pages 11–53. Wiley, New York.

Henian Chen, Patricia Cohen, and Sophie Chen. 2010. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, 39(4):860–864.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

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.

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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Embick and Alec Marantz. 2008. Architecture and blocking. In *Linguistic Inquiry*, volume 39, pages 1–53. MIT Press, Cambridge, MA.

Kyle Gorman. 2023. Notes on morphological defectivity. Lingbuzz preprint. Handout from an invited talk given at the University of Surrey.

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.

Christer Johansson. 1999. Learning what cannot be by failing expectations. *Nordic Journal of Linguistics*, 22:61–76.

Christo Kirov, John Sylak-Glassman, Ryan Que, David Yarowsky, Jason Eisner, and Ryan Cotterell. 2018. Universal morphological inflection generation using a multilingual dataset. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 52–62.

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.

Renato Oniga and Norma Shifano. 2014. *Latin: A Linguistic Introduction*. Oxford University Press, Oxford.

Cemil Orhan Orgun and Ronald L. Sprouse. 1999. From MPARSE to CONTROL: Deriving ungrammaticality. *Phonology*, 16(2):191–224.

Marco Passarotti, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, and Paolo Ruffolo. 2019. The LiLa knowledge base of linguistic resources and NLP tools for Latin. In *Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019)*, Leipzig, Germany.

Andrea D. Sims. 2006. *Minding the Gaps: Inflectional Defectiveness in a Paradigmatic Theory*. Ph.D. thesis, The Ohio State University.

Rachele Sprugnoli, Margherita Fantoli, Flavio Massimiliano Cecchini, and Marco Passarotti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, pages 183–189.

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).

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzm'an, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Daniel Yakubov. 2024. How do we learn what we cannot say? Master's thesis, CUNY Graduate Center.