The UD-NewsCrawl Treebank: Reflections and Challenges from a Large-scale Tagalog Syntactic Annotation Project

Angelina A. Aquino[2,4]*   Lester James V. Miranda[1]*   Elsie Marie T. Or[3]*
[1]Allen Institute for AI [2]Charles Darwin University
[3]Department of Linguistics, University of the Philippines Diliman
[4]Electrical and Electronics Engineering Institute, University of the Philippines Diliman
angelina.aquino@cdu.edu.au,ljvmiranda@gmail.com,etor@up.edu.ph

🤗 **Dataset** hf.co/datasets/UD-Filipino/UD_Tagalog-NewsCrawl

Abstract

This paper presents UD-NewsCrawl, the largest Tagalog treebank to date, containing 15.6k trees manually annotated according to the Universal Dependencies framework. We detail our treebank development process, including data collection, pre-processing, manual annotation, and quality assurance procedures. We provide baseline evaluations using multiple transformer-based models to assess the performance of state-of-the-art dependency parsers on Tagalog. We also highlight challenges in the syntactic analysis of Tagalog given its distinctive grammatical properties, and discuss its implications for the annotation of this treebank. We anticipate that UD-NewsCrawl and our baseline model implementations will serve as valuable resources for advancing computational linguistics research in underrepresented languages like Tagalog.

1 Introduction

The Philippine archipelago is home to over 170 languages. Among these, Tagalog is the most well-known and widely spoken, as the native language of the country's capital region and a "*de facto* national working language" (Eberhard et al., 2024) of education, governance, trade, and many other domains of communication. The native speakers of Tagalog reside mostly in the southwestern regions of Luzon, the Philippines' largest island, but Tagalog has also been adopted as a second language in many other parts of the country (Himmelmann, 2005). Roughly 25 percent of Filipinos (out of a national population of over 100 million) identify as being of Tagalog ethnicity, while 40 percent of households in the country use it as one of their home languages (Philippine Statistics Authority). Large Tagalog-speaking populations have also been recorded in many other countries, including over 1.3 million Tagalog speakers in the United States and over 700,000 speakers each in Canada and Saudi Arabia (Eberhard et al., 2024).

Despite the widespread use of Tagalog, large-scale NLP resources for the language remain sparse (Joshi et al., 2020; Cruz and Cheng, 2022; Miranda, 2023b). At present, only two Tagalog treebanks, annotated using the Universal Dependencies (UD) framework (de Marneffe et al., 2021; Nivre et al., 2020)—a standardized annotation scheme for linguistic data—are available for public use: UD-Ugnayan (Aquino and de Leon, 2020) and UD-TRG (*Tagalog Reference Grammar*, Samson, 2018). Together, these treebanks contain just 149 sentences, which are dwarfed by treebanks of high-resource languages and are insufficient for training robust dependency parsers. Furthermore, Tagalog treebanks lag behind those of its Southeast Asian neighbors in terms of size, such as Indonesian (7.6k sentences), Vietnamese (3.4k), and Thai (1.0k).[1] This underscores a need for more comprehensive computational resources to support the development of Tagalog language technologies.

In this work, we present UD-NewsCrawl, the largest Tagalog treebank to date, consisting of 15,619 sentences and manually annotated in accordance with the UD framework. The treebank is composed of text sourced from the Leipzig Tagalog NewsCrawl corpus.[2] The text was annotated by native speakers with domain knowledge in linguistics, and the annotations were verified through manual inspection and semi-automated quality control. Furthermore, we introduce transformer-based parsers on UD-NewsCrawl which serve as baselines for further research into Tagalog dependency parsing. The UD-NewsCrawl treebank and baseline models are available publicly.

*Equal contributions.

[1] https://universaldependencies.org/

[2] https://corpora.uni-leipzig.de/en?corpusId=tgl_newscrawl_2011#tgl

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7219–7239
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Universal Dependencies (UD) and Tagalog treebanks. The Universal Dependencies framework aims to provide a consistent annotation schema for parts-of-speech (POS) tagging, morphological features, and dependency relations across languages. UD falls under the class of dependency grammars, and follows a principle of content word primacy (Muischnek et al., 2016; Dirix et al., 2017), i.e. syntactic relations in UD primarily hold between content words, and syntactic structures are typically headed by content words, with function words being dependencies thereof.

As of writing, there are two Tagalog treebanks which apply this framework: UD-Ugnayan (Aquino and de Leon, 2020) and UD-TRG (Samson, 2018). The former contains 94 sentences from educational fiction and non-fiction text drawn from the Philippine Department of Education's Learning Resource Portal, while the latter contains 55 sentences from Tagalog grammar books (Schachter and Otanes, 1972; de Vos, 2010). However, due to their size, the UD dataset guidelines recommend these treebanks to be treated as test data (and to use 10-fold cross-validation for evaluation if one wishes to train on the dataset), and this limits scalable development of NLP models.

Tagalog dependency parsing. Several works have used UD-Ugnayan and UD-TRG for automated dependency parsing. For example, Aquino and de Leon (2020) showed that even a small treebank like UD-Ugnayan can be used to develop dependency parsers, trained using UDPipe (Straka et al., 2016) and Stanza (Qi et al., 2020), that are competitive with cross-lingual or multilingual parsers. They have also shown that decent tokenization and tagging performance can be achieved using alternative language resources and data augmentation (Aquino and de Leon, 2022). On the other hand, Miranda (2023a) trained a dependency parser by combining both treebanks and using the spaCy framework (Honnibal et al., 2020) based on pre-trained RoBERTa embeddings (Conneau et al., 2020) . Due to the absence of a canonical train, development, and test split, prior works resort to different evaluation paradigms such as k-fold cross-validation that may not be directly comparable, hampering effective benchmarking and consistent assessment of parser performance across studies.

3 Background

3.1 The Tagalog language

Tagalog, an Austronesian language belonging to the Western Malayo-Polynesian branch of the language family (Blust, 1991), occupies a prominent position within the field of linguistics, having been documented since the Spanish occupation of the Philippines in the 16th century, and playing a part in the development of the American structuralist movement the beginning of the 20th century (Javier and Or, 2022). It has become the most well-known representative of the so-called "Philippine-type languages"—a designation for a group of Austronesian languages that share a distinctive grammatical "voice marking" system (Himmelmann, 2005; Reid, 2005) which we discuss in the next section.

Tagalog has a non-configurational phrase structure (Kroeger, 1993) wherein sentences are canonically predicate-initial and sentence arguments following the predicate have flexible order and can also be fronted using the *ay* inversion marker. Tagalog lacks a copula, and thereby permits noun phrases, prepositional phrases, and adjectival forms to be sentence predicates (Reid, 2005); specificational and identificational clauses in Tagalog are formed by a simple juxtaposition of the subject and complement. The language is also noted for its productive affixation, which allows lexical terms from typologically dissimilar languages like English to be easily encoded in its morphosyntax (Tangco and Nolasco, 2002), and its varied forms and functions of reduplication (Blake, 1917).

3.2 UD annotation of Tagalog

Despite extensive linguistic scholarship on Tagalog, some features of Tagalog grammar have been the topic of continued debate among linguists (Javier and Or, 2022). In annotating the UD-NewsCrawl treebank, we made decisions which reveal our positions in some of these debates. Here we discuss our approaches to two major points of contention in the syntactic analysis of Tagalog; several other annotation choices dealing with specifics of the UD framework are outlined in Appendix B. The possible typological implications of these choices are outside the scope of this paper but receive a more thorough treatment in Bardají et al. (2024).

Voice marking system. One debate in Tagalog grammar is related to its voice marking system and

Nag-bigay	***ang***	***lalaki***	*ng*	*bulaklak*	*sa*	*babae*
AV.PRF-give	NOM	man	GEN	flower	LOC	woman
B<in>igay	*ng*	*lalaki*	***ang***	***bulaklak***	*sa*	*babae*
<PV.PRF>give	GEN	man	NOM	flower	LOC	woman
B<in>igy-an	*ng*	*lalaki*	*ng*	*bulaklak*	***ang***	***babae***
<PRF>give-LV	GEN	man	GEN	flower	NOM	woman

Figure 1: Example sentences illustrating features of Tagalog voice marking under a symmetrical voice analysis: each sentence has a different **subject** (preceded by the ***ang*** marker) and a different verbal affix denoting the thematic role (AV = agent, PV = patient, LV = locative) of the subject, but all three examples are pragmatically equivalent to the English sentence "The man gave flowers to the woman."

assignment of grammatical relations (Cubar, 1975; Schachter, 1976; Rafael, 2016). Tagalog and other "Philippine-type languages" are said to have a distinctive voice marking system, wherein essentially any type of clausal argument—which may be a noun phrase (NP) referring to an entity or object that performs the thematic role of agent, patient, location, beneficiary, or instrument—can be marked as subject depending on the voice affix attached to the predicate head. In Figure 1, the NP preceded by the marker *ang* is generally viewed as the subject in the sentence and agrees with the voice marking on the predicate head, while the markers *ng* and *sa* precede non-subject arguments and adjuncts. These markers have respective counterparts for marking proper nouns.

For this treebank annotation project, we adopted a **symmetrical voice** view of Tagalog voice marking. Following this view, we consider all three sentences in Figure 1 as "unmarked" or basic transitive sentences, as such languages do not show a preference for the agent argument to be the subject, unlike in Indo-European languages like English (Foley, 2008; Riesberg et al., 2019).

In Figure 1 therefore, while the verbs in each sentence exhibit different voice markings and the subject carry different thematic roles, the examples could all be translated as the equivalent of the English sentence "The man gave flowers to the woman," given that no process of agent demotion takes place in Tagalog constructions where the agent is not the focused argument, which is contrary to what would happen in English passive voice constructions. However, it should also be noted that a speaker's choice of voice is usually affected by different factors, which could include definiteness (Himmelmann, 2005), specificity (Rackowski and Richards, 2005), and topicality in discourse (Carrier-Duncan, 1985).

The type of voice marking used in a clause is indicated at the morphological features level of our UD annotation scheme, which is similar to how they are marked in the UD-TRG treebank. The following voice types were marked in UD-NEWSCRAWL: (1) `Act`, which is the label UD assigns to both the Indo-European active voice and the actor-focus voice of Austronesian languages; (2) `Pass`, which is short for passive but which also applies to the Austronesian patient-focus voice, (3) `Bfoc` for beneficiary-focus voice, (4) `Lfoc` for location-focus voice, and (5) `Cau` for causative forms which UD classifies as a voice category.

In indicating the dependency relation between the predicate head and its core arguments, we met theoretical issues regarding the working definitions of certain categories in the UD framework. For example, the `nsubj` relation is defined in terms of the semantic role carried by the subject NP and favors grammatical systems such as those seen in Indo-European languages like English where the agent is usually seen as the more privileged argument. While such UD labels can be edited to a certain extent in the language-specific documentation to better suit the grammar of a language, incompatibilities such as in the example provided reveals how English has become the *de facto* standard for understanding how languages work.

Categorization of Tagalog roots. The lexical categories of Tagalog roots and the question of whether or not they should be considered precategorial prior to affixation have also been a topic of debate among linguists (see De Guzman, 1978; Himmelmann, 1991, 2005; Kaufman, 2009). Himmelmann (2007), for example, observed that in several Tagalog dictionaries, many roots that could be presumed verbal are glossed with English nouns or adjectives, such as 'gift' for *bigay* and 'surpassed,

	# Sents / Tokens			# Tags	
Treebank	Train	Dev	Test	UPOS	DEPREL
UD-NewsCrawl	12.4k / 286.9k	1.56k / 37.0k	1.56k / 36.9k	17	39
UD-TRG (Samson, 2018)	–	–	128 / 734	13	26
UD-Ugnayan (Aquino and de Leon, 2020)	–	–	94 / 1.01k	14	24

Table 1: Dataset statistics for UD-NewsCrawl and its comparison to existing Tagalog treebanks in the Universal Dependencies (UD) framework. The breakdown of tags per split is found in Appendix A.

defeated' for *daig*, even though these can be inflected with verb affixes, thus they can alternatively be glossed as 'to give' and 'to surpass or defeat', respectively.

The POS designation of Tagalog roots in UD-NewsCrawl was determined by the morphological structure of the word. Thus, if a root is affixed with a nominal affix, such as *pag-* in the word *paglakad* (manner of walking), then it is labeled as NOUN. However, if a verbal affix is attached to it, such as *mag-* in *maglakad* 'to walk', then it is labeled as VERB. Meanwhile, the POS of words that are unaffixed or appear in their bare forms are usually determined by their syntactic distribution. Otherwise, the Tagalog-English dictionary of Leo James English was consulted if the POS of a word remains ambiguous to the manual annotator.

4 The UD-NewsCrawl treebank

UD-NewsCrawl contains 15.6k sentences (360.8k tokens) manually annotated by Tagalog native speakers. Individual words in the dataset underwent lemmatization and POS tagging. Dependency relations of words in each sentence were identified, and select morphological features were also annotated. Dataset details can be found in Table 1. In this section, we describe the context of the annotation project, the nature of text included in the corpus, and our procedures for annotation and quality control.

4.1 Project context

UD-NewsCrawl was originally developed to fulfill a different linguistic analysis objective, i.e., to investigate cross-linguistic variation in the distribution of lexical information, especially on languages like Tagalog that exhibit different degrees of formal distinction between their syntactic categories (§3.2). As a result, there were annotation decisions incompatible with the original UD guidelines (as seen in Appendix B). After creating the initial version of the treebank, we perform post-processing to ensure that it meets current UD standards (§4.4).

We recruited a total of 15 annotators to manually label UD-NewsCrawl using the WebAnno 3.6.4 annotation platform. All annotators are native speakers of Tagalog, with most annotators having an undergraduate to postgraduate level of linguistics education. Annotators were given prior training on linguistic concepts involved in UD syntactic analysis, as well as the use of the WebAnno platform. They were then compensated at a rate of ₱2500 (roughly €40) for each set of 100 sentences annotated, with an estimated workload of 20 hours per set. The annotation project spanned 16 months, including planning and actual annotation, while the quality control process lasted for a year.

4.2 Text corpus properties

The text included in the treebank is sourced from the Leipzig Tagalog NewsCrawl 2011 corpus, which consists of material collected from Tagalog news sites (Quasthoff and Richter, 1998; Goldhahn et al., 2012). Sample sentences from the corpus can be found in Appendix F, with most sentences written in the declarative and more formal register of Tagalog. Some use of informal language can also be found in the corpus, particularly when texts quote a person's speech. We also find several instances of code-switching in the texts, reflecting the bilingual nature of many Tagalog speakers.

The data collection team automated the sentence tokenization of texts in the corpus and selected a subset of 15,000 sentences. These were then divided into 150 files of 100 sentences each in order of increasing sentence length, with an approximately even distribution of sentence lengths across files. This data preparation process distributed the annotation difficulty across files, and allowed an annotator working on one file to start with shorter, simpler sentences then progress to longer, more challenging sentences as their proficiency in the annotation workflow increased. However, the data preparation process also resulted in

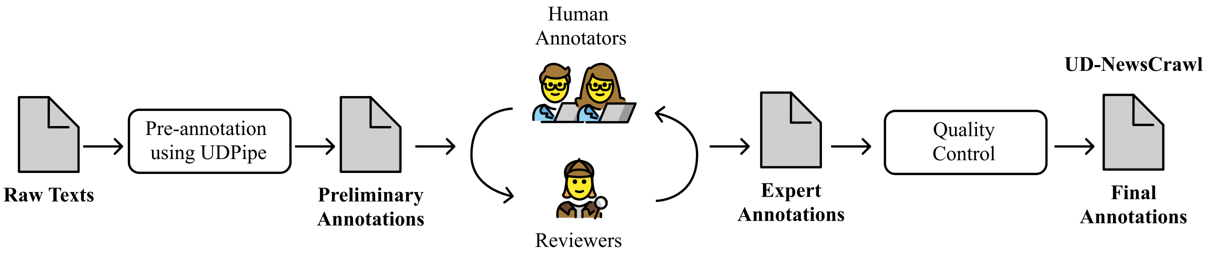


Figure 2: Annotation workflow for UD-NewsCrawl.

sentences being isolated from their original context, and some errors in the automated sentence tokenization produced fragments or run-on sentences instead, which made the subsequent annotation more challenging.

4.3 Annotation procedure

An initial set of annotation guidelines was prepared based on the UD framework and the project's analysis objectives. We then employed an iterative annotation workflow (c.f. Fort, 2016) for UD-NewsCrawl, as seen in Figure 2, which enabled us to modify and expand on the annotation guidelines as we encountered cases in the texts which were not previously covered. We highlight aspects of our annotation guidelines in Appendix B.

Pre-annotation. Prior to human annotation, POS tags and lemmas were annotated using a UDPipe model trained on the UD-Ugnayan treebank, resulting in a set of **preliminary annotations**. This reduced the cognitive load of annotators who only needed to validate the pre-annotated labels and correct as needed instead of labeling from scratch.

We opted not to pre-annotate morphological features and dependency relations, since the higher complexity of word forms and sentence structures found in the NewsCrawl corpus resulted in a high error rate for these annotations from models trained on existing Tagalog treebanks with simpler texts.

Human annotation. Annotators were tasked to first validate and correct tokenization, especially the splitting of linkers and contractions. Once the tokenization was validated, annotators proceeded to label POS tags, dependency relations, and morphological features following the working annotation guidelines. Because of differences in annotators' competency in linguistic annotation, the annotation workload for some files was divided between two annotators: less experienced annotators handled tokenization and POS tagging while more experienced annotators were tasked with labeling dependency relations and morphological features. Finally, we encouraged annotators to keep a set of notes to document edge cases and unusual examples for further deliberation.

Verification and re-annotation. A senior linguist independently reviewed the annotators' work to validate decisions and identify potential inconsistencies. We discussed and resolved difficult annotation cases in regular meetings, and updated the annotation guidelines accordingly. Cases requiring revision were re-annotated following the updated annotation guidelines. This step helped maintain quality and consistency across the dataset. This cycle of annotation, verification, and re-annotation resulted in a set of **expert annotations** that we send for quality control.

4.4 Treebank quality-control

To ensure that the expert annotations remained consistent throughout the treebank, we employed both semi-automated and manual post-annotation workflows. First, we trained a silver-standard parsing model on our existing annotations using spaCy (Honnibal et al., 2020) and identified instances where the model's morphological annotations and dependency parsing relations disagree with human annotations. This approach is based on the premise that a silver-standard model can learn global patterns from the training labels, even when trained on partially incorrect data (Tedeschi et al., 2021; Zhang et al., 2022; Wang et al., 2024). These cases were prioritized for review, as disagreements often indicated potential inconsistencies or errors. Second, we conducted manual quality checks through random sampling of sentences, comparing them against established UD treebanks using the UD official validator[3] to ensure adherence to UD guidelines and consistency. We describe the results of

[3] https://github.com/UniversalDependencies/tools

Pipeline	**Lemm.**	**UPOS**	**Morph.**		**Dep. Parsing**	
Feature representation	*Acc*	*Acc*	*Acc*	*F1-score*	*UAS*	*LAS*
No embeddings	89.5±1.1	89.7±0.8	94.3±0.8	92.8±0.4	82.4±1.0	75.4±0.8
Word embeddings						
fastText (Bojanowski et al., 2017)	89.8±0.8	90.3±0.3	94.9±0.3	93.9±0.2	83.1±0.5	76.2±0.5
Multi hash embeddings (Miranda et al., 2022)	90.3±0.8	90.9±0.1	95.4±0.1	94.6±0.1	83.9±0.5	77.4±0.6
Context-sensitive vectors						
mDeBERTa-v3, base (He et al., 2021)	90.6±1.3	91.3±0.8	95.2±0.7	94.7±0.7	85.1±0.8	79.0±0.6
RoBERTa-Tagalog, large (Cruz and Cheng, 2022)	90.4±0.9	**91.7±0.7**	**95.6±0.8**	**95.1±0.6**	86.4±0.7	80.6±0.6
XLM-RoBERTa, large (Conneau et al., 2020)	**91.0±0.9**	91.5±0.9	95.4±0.8	**95.1±0.6**	**86.9±0.0**	**81.0±0.1**

Table 2: Test set performance on various linguistic tasks using the UD-NEWSCRAWL given different feature representations. We report the average of three runs and their standard deviation. Full results can be found in Appendix D.3.

this error analysis in Appendix I. After quality control, we obtain a set of **final annotations** which forms the basis for UD-NEWSCRAWL and succeeding experiments.

5 Baseline Models for UD-NEWSCRAWL

Set-up. To establish baseline performance for UD-NEWSCRAWL, we trained several multi-task models that accommodate different compute requirements using the spaCy framework (Honnibal et al., 2020). Each model consists of the following components (full description in Appendix C):

- **Lemmatizer**: employs an edit-tree recursive algorithm based on Müller et al. (2015). We train a convolutional network with a softmax layer to predict the best edit-tree for a token.
- **Morphological and UPOS tagger**: treats morphological annotation and POS tagging as a multilabel classification problem and implements a softmax layer to predict scores given a token.
- **Dependency parser**: uses a variant of the non-monotonic arc-eager transition system as described in Honnibal and Johnson (2015).

In order to investigate how different feature representations affect model performance, we trained these components using several feature representations: (1) fastText word embeddings (Bojanowski et al., 2017), (2) spaCy's multi hash embeddings as described in Miranda et al. (2022), (3) monolingual context-sensitive vectors using RoBERTa Tagalog (Cruz and Cheng, 2022), and (4) multilingual context-sensitive vectors using XLM-RoBERTa (Conneau et al., 2020) and mDeBERTa-v3 (He et al., 2021). For all context-sensitive vectors, we use the `large` variant of the pretrained models except for mDeBERTa-v3, which is not available. The full training hyperparameters can be found in Appendix D. Finally, we evaluated each component in its corresponding linguistic task. We report the accuracy for both lemmatization and POS tagging tasks and the macro F1-score for the morphological annotation task. In addition, we also report the unlabeled and labeled attachment scores (UAS / LAS) for the dependency parsing task.

Results. Table 2 presents the performance of various baseline models trained on UD-NEWSCRAWL. The model trained on multilingual XLM-RoBERTa context-sensitive vectors achieves the best performance in most tasks, with 1–2% (absolute) improvement from the "No embeddings" baseline. This performance is closely followed by monolingual RoBERTa, which was trained specifically on Tagalog texts. We hypothesize that XLM-RoBERTa leverages cross-lingual transfer (Artetxe et al., 2019), allowing it to perform at par with a fully-monolingual model.

6 Analysis

We perform several analyses to evaluate the quality of expert annotations (§6.1), the generalization of UD-NEWSCRAWL to other Tagalog treebanks (§6.2), and the type of content that exists within the treebank (§6.3).

6.1 Quality analysis

Set-up. During the quality control process (§4.4), we developed a silver-standard model to compare its annotations against expert annotations. We focused on identifying and correcting sentences where discrepancies occurred. To evaluate the level

Metric	Cohen's κ	% Corrected
UPOS	0.75	15%
Dep. Rel.	0.68	20%
Morph.	0.70	22%

Table 3: Initial disagreement between the silver-standard model and original annotations, and the proportion of disagreed sentences corrected.

of agreement, we used Cohen's κ and calculated the proportion of sentences corrected out of all the cases with disagreements.

Results. Table 3 shows the initial disagreement, as measured by Cohen's κ between the silver-standard model and the original annotations, together with the proportion of the disagreed sentences that we corrected. The results indicate moderate agreement between the silver-standard model and the original annotations (McHugh, 2012). These results imply that while the annotations are fairly consistent, there is room for improvement, especially in dependency relations and morphological features. The percentage of corrected sentences shows that a small portion of the annotations required adjustments, highlighting areas where the silver-standard model's predictions diverged from human annotations.

6.2 Cross-treebank generalization on UD-TRG and UD-Ugnayan

Set-up. In order to understand how well the parsers trained on UD-NewsCrawl generalize to other datasets, we used the best performing model—i.e., with components trained on context-sensitive vectors from XLM-RoBERTa—and evaluate it on the UD-TRG and UD-Ugnayan treebanks. We also compared against the best reported results from the literature for both treebanks such as Dehouck and Denis (2019)'s phylogenic tree approach and Kondratyuk and Straka (2019)'s UDify for UD-TRG, and Aquino and de Leon (2020)'s cross-lingual approach for UD-Ugnayan.

Results. Our results in Table 4 suggest that UD-NewsCrawl effectively captures the linguistic structures and patterns of Tagalog, making it a robust resource for dependency parsing. However, we note that the results are severely affected by the small size of the other treebanks, which hinders a proper comparison. Despite this, the models achieve state-of-the-art performance by surpassing previously reported benchmarks on both treebanks. The cross-treebank generalization indicates that despite potential differences in domain, annotation style, or text genre, the syntactic patterns encoded in these models are potentially transferrable across various contexts.

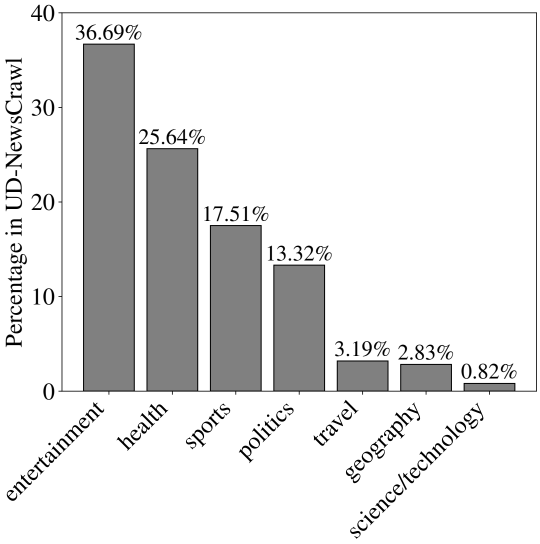


Figure 3: Topic distribution of UD-NewsCrawl using categories from SIB-200 (Adelani et al., 2024).

6.3 Topics in UD-NewsCrawl

Set-up. In order to identify the topics present in UD-NewsCrawl, we performed zero-shot classification using Llama-3.1-8B-Instruct (Dubey et al., 2024) given a set of topics defined in SIB-200 (Adelani et al., 2024) (prompt template can be found in Appendix H), then manually evaluated the results. Figure 3 shows the topic distribution for UD-NewsCrawl. Examples for each topic can be found in Appendix F. In addition, we also perform a more fine-grained topic analysis as seen in Appendix G

Results. Our analysis of the topic distribution reveals a clear dominance of entertainment and health-related content, constituting over 60% of the dataset. This skewed distribution highlights potential collection biases in web-crawled news data, which may impact downstream applications relying on this dataset for training or evaluation.

7 Discussion

Syntactic annotation in a time of LLMs. One key motivation for training dependency parsers (and consequently, creating treebanks) has been their downstream utility in NLP applications such

Treebank	Lemm. *Acc*	UPOS *Acc*	Morph. *Acc*	 *F1-score*	Dep. Parsing *UAS*	 *LAS*
UD-TRG						
Ours	**81.1**±0.4	**78.3**±0.6	**73.4**±0.4	**78.5**±0.4	**95.5**±0.2	**68.5**±0.4
Dehouck and Denis (2019)	–	–	–	–	70.9	50.4
Kondratyuk and Straka (2019)	75.0	61.6	35.3	–	64.7	39.4
UD-Ugnayan						
Ours	83.3±0.4	**82.3**±0.4	–	–	**82.8**±0.4	**60.8**±0.4
Aquino and de Leon (2020)	**85.5**	80.5	–	–	63.5	55.4

Table 4: Performance comparison of our best performing pipeline (**Ours**), i.e., context-sensitive vectors from XLM-RoBERTa trained on UD-NewsCrawl, on the UD-TRG and UD-Ugnayan treebanks against previous reported results in literature. Reporting the average of three runs and the standard deviation for our results.

as question answering, grammar rule extraction, and semantic role labeling (Liang et al., 2011; Berant et al., 2013; Herrera et al., 2024, *inter alia*). However, the emergence of large language models (LLMs) that can directly perform many of these tasks brings into question the continued relevance of labor-intensive treebank creation.

We argue that syntactic annotations remain valuable not just as training data, but as **explicit, interpretable representations of linguistic structure** that enable detailed analysis of language phenomena, hypothesis testing about grammar, and evaluation of model capabilities in ways that end-task performance alone cannot capture. As noted by Lappin (2024), LLMs are biased towards certain language expressions by both their architecture and represented distributions in the training data (Ryan et al., 2024; Bhatt and Diaz, 2024, *inter alia*). Additionally, LLMs are opaque as we cannot fully explain their representations of language nor reliably modify them to produce set outcomes.

Challenging the universality of UD. While Universal Dependencies aims to provide a cross-linguistically consistent annotation scheme, several linguistic phenomena challenge its universality (Osborne and Gerdes, 2019; Kanayama and Iwamoto, 2020). For example, in UD-NewsCrawl, the `nsubj` relation is defined in terms of subject NP semantic roles typical of Indo-European languages. Although UD provides language-specific documentation, English (and Euro-western linguistic frameworks) remains the default reference point for analyzing syntactic structures.

We highlight the importance for NLP practitioners to reflect on the implicit biases inherent in current linguistic frameworks and model architectures, which are often initialized and optimized based on globally dominant languages. This reflection is crucial for developing more equitable and effective NLP tools that respect and accommodate linguistic diversity.

8 Conclusion

In this work, we introduced UD-NewsCrawl, the largest Tagalog treebank to date, consisting of 15,619 sentences manually annotated according to the Universal Dependencies framework. We described our development process for UD-NewsCrawl, which involved multiple stages of manual and semi-automated annotation and verification to ensure the quality of the treebank. We also discussed several linguistic challenges specific to Tagalog which required careful consideration during the annotation process. Our baseline evaluations and experiments using transformer-based models trained on UD-NewsCrawl demonstrate the robustness and generalizability of the dataset, highlighting its utility in different areas of computational linguistics research and development.

We anticipate that UD-NewsCrawl will serve as a valuable resource for researchers and practitioners working on Tagalog NLP, enabling the development of more accurate and robust language technologies. We hope that insights gained from this project can also inform future efforts to create syntactic resources for other low-resource languages, contributing to a more inclusive and diverse NLP landscape.

Author contributions

Or led the annotation project, with Aquino as the technical consultant for UD annotation. Miranda assisted with the treebank quality-control and on training baselines from the treebank. All authors contributed in drafting and writing the manuscript.

Limitations

Domain and topic bias The corpus is primarily composed of news articles, which may introduce domain-specific biases. This could limit the generalizability of models trained on UD-NewsCrawl to other domains, such as conversational or literary texts. In addition, the corpus is skewed towards entertainment and health-related content, which may not fully represent the diversity of topics in Tagalog. This could impact the performance of models in downstream applications that require broader topic coverage. Finally, the timeframe of the texts extends until 2011, potentially missing more recent developments in language use and contemporary topics. This temporal limitation may affect the model's ability to process modern terminology and current cultural references, particularly in rapidly evolving domains such as technology and social media.

Annotation challenges Despite rigorous quality control, the annotation process faced challenges due to the context of the project. For example, a portion of the annotation work were done during the height of the COVID-19 pandemic (from January 2021 to April 2022), making it difficult for some annotators to stay throughout the project. Some inconsistencies may still exist despite quality-control, especially in edge cases.

Ethics Statement

The development of UD-NewsCrawl involved the manual annotation of texts by native Tagalog speakers, who were compensated for their work. We ensured that the annotators were treated fairly and that their contributions were acknowledged. The data used in the corpus was sourced from publicly available news articles, and we adhered to ethical guidelines regarding data usage and privacy. However, it is important to consider the potential ethical implications of using web-crawled data, particularly in terms of copyright and the representation of diverse voices. While the corpus reflects the bilingual nature of many Tagalog speakers, it may not fully capture the linguistic diversity of the Philippines, which is home to over 170 languages. Future efforts should aim to include a more diverse range of texts and dialects to ensure broader representation. Finally, the creation of UD-NewsCrawl is intended to support the development of NLP applications for Tagalog, a language that has been historically underrepresented in computational linguistics. By providing this resource, we hope to contribute to the preservation and promotion of Tagalog in the digital age, while also encouraging similar efforts for other low-resource languages.

Acknowledgements

The treebank was originally created for the Deutsche Forschungsgemeinschaft (DFG)-funded project entitled "Information distribution and language structure - correlation of grammatical expressions of the noun/verb distinction and lexical information content in Tagalog, Indonesian and German" which was headed by Gerhard Heyer from Leipzig University and Nikolaus P. Himmelmann from the University of Cologne. The authors would also like to thank all annotators from the Linguistics Department of the University of the Philippines - Diliman, who helped in the project: Patricia Anne Asuncion, Paola Ellaine Luzon, Jenard Tricano, Mary Dianne Jamindang, Michael Wilson Rosero, Jim Bagano, Yeddah Joy Piedad, Farah Cunanan, Calen Manzano, Aien Gengania, Prince Heinreich Omang, Noah Cruz, Leila Ysabelle Suarez, Orlyn Joyce Esquivel, and Andre Magpantay. Finally, we would also like to thank the reviewers during the February ARR cycle for their thoughtful comments and feedback that helped improve this paper.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Angelina Aquino and Franz de Leon. 2020. Parsing in the absence of related languages: Evaluating low-resource dependency parsers on Tagalog. In *Proceedings of the Fourth Workshop on Universal Depen-*

dencies (UDW 2020), pages 8–15, Barcelona, Spain (Online). Association for Computational Linguistics.

Angelina Aquino and Franz de Leon. 2022. Zero-shot and few-shot approaches for tokenization, tagging, and dependency parsing of Tagalog text. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 190–202, Manila, Philippines. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Maria Bardají, Elsie Or, Angelina Aquino, and Nikolaus Himmelmann. 2024. The challenges of symmetrical voice languages for universal dependencies. In *Proceedings of the 15th International Conference of the Association for Linguistic Typology*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

Frank R Blake. 1917. Reduplication in tagalog. *The American journal of philology*, 38(4):425–431.

Robert Blust. 1991. The Greater Central Philippines Hypothesis. *Oceanic Linguistics*, 30(2):73–129.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jill Carrier-Duncan. 1985. Linking of thematic roles in derivational word formation. *Linguistic Inquiry*, 16(1):1–34.

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.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

Ernesto H. Cubar. 1975. *Topicalization and Some Related Processes in Philippine Languages*. University of the Philippines Department of Linguistics. Reprinted in The Archive Classics (2019).

V. P. De Guzman. 1978. *Syntactic Derivation of Tagalog Verbs*. University of Hawaii Press, Honolulu.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Fiona de Vos. 2010. *Essential Tagalog Grammar: A Reference for Learners of Tagalog*. Learning Tagalog.

Mathieu Dehouck and Pascal Denis. 2019. Phylogenic multi-lingual dependency parsing. 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 192–203, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Dirix, Liesbeth Augustinus, Daniel van Niekerk, and Frank Van Eynde. 2017. Universal Dependencies for Afrikaans. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 38–47, Gothenburg, Sweden. Association for Computational Linguistics.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath. 2013. WALS Online (v2020.4). Data set.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, 27 edition. SIL International, Dallas, Texas.

Leo James English. 1986. *Tagalog-English dictionary*. Congregation of the Most Holy Redeemer Manila.

William Foley. 2008. The place of philippine languages in a typology of voice systems. In P. K. Austin and S. Musgraves, editors, *Voice and grammatical relations in Austronesian languages*, pages 22–44. CSLI Publications.

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Santiago Herrera, Caio Corro, and Sylvain Kahane. 2024. Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125, Torino, Italia. ELRA and ICCL.

Nikolaus P. Himmelmann. 1991. *The Philippine Challenge to Universal Grammar*. Number N.F. 15 in Arbeitspapier / Institut für Sprachwissenschaft, Universität Köln.

Nikolaus P. Himmelmann. 2005. Tagalog. In K. Alexander Adelaar and Nikolaus P. Himmelmann, editors, *The Austronesian Languages of Asia and Madagascar*, 1st edition, page 27. Routledge.

Nikolaus P. Himmelmann. 2007. Lexical categories and voice in tagalog. In P. K. Austin and S. Musgrave, editors, *Voice and grammatical relations in Austronesian languages*, pages 247–293.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jem R Javier and Elsie Marie T Or. 2022. Tagalog linguistics: Historical development and theoretical trends. In *The Routledge Handbook of Asian Linguistics*, pages 33–46. Routledge.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Hiroshi Kanayama and Ran Iwamoto. 2020. How universal are Universal Dependencies? exploiting syntax for multilingual clause-level sentiment detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4063–4073, Marseille, France. European Language Resources Association.

Daniel Kaufman. 2009. Austronesian nominalism and its consequences: A tagalog case study. *Theoretical Linguistics*, 35(1):1–49.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Paul Kroeger. 1993. *Phrase structure and grammatical relations in Tagalog*. Center for the Study of Language (CSLI).

Shalom Lappin. 2024. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*, 33(1):9–20.

Percy Liang, Michael Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 590–599, Portland, Oregon, USA. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Lester James Miranda. 2023a. calamanCy: A Tagalog natural language processing toolkit. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 1–7, Singapore. Association for Computational Linguistics.

Lester James Miranda. 2023b. Developing a named entity recognition dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Lester James Miranda, Ákos Kádár, Adriane Boyd, Sofie Van Landeghem, Anders Søgaard, and Matthew Honnibal. 2022. Multi hash embeddings in spacy. *arXiv preprint arXiv:2212.09255*.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian dependency treebank: from constraint grammar tagset to Universal Dependencies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1558–1565, Portorož, Slovenia. European Language Resources Association (ELRA).

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of universal dependencies (ud). *Glossa: a journal of general linguistics (2016-2021)*.

Philippine Statistics Authority. 2020 census of population and housing report no. 21 - demographic and housing characteristics (non-sample variables). Retrieved from https://library.psa.gov.ph/.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Uwe Quasthoff and Matthias Richter. 1998. *Projekt Der Deutsche Wortschatz*. na.

Andrea Rackowski and Norvin Richards. 2005. Phrase edge and extraction: A tagalog case study. *Linguistic Inquiry*, 36(4):565–599.

Ria P. Rafael. 2016. Sinong pasimuno? paggamit ng subject at topic sa pag-aaral ng wika' (who's the initiator? the use of subject and topic in language studies). *Daluyan: Journal ng Wikang Filipino*, 22(1/2):165–180.

Lawrence Reid. 2005. Tagalog and philippine languages. In Philipp Skutch, editor, *Encyclopedia of Linguistics*. Routledge, New York.

Sonja Riesberg, Kurt Malcher, and Nikolaus P. Himmelmann. 2019. How universal is agent-first? evidence from symmetrical voice languages. *Language*, 95(3):523–561.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Dawn Samson. 2018. A treebank prototype of Tagalog. Bachelor's thesis, University of Tübingen, Germany.

Paul Schachter. 1976. The subject in philippine languages: Topic, actor, actor-topic, or none of the above. In Charles Li, editor, *Subject and Topic*, pages 491–518. Academic Press, New York.

Paul Schachter and Fe T. Otanes. 1972. *Tagalog Reference Grammar*. University of California Press.

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).

Roberto D Tangco and Ricardo Ma Nolasco. 2002. 'taglish'verbs: How english loanwords make it into the philippine languages. In *Tenth Annual Meeting of the Southeast Asian Linguistics Society*, pages 391–406.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianwei Wang, Tianyin Wang, and Ziqian Zeng. 2024. On the use of silver standard data for zero-shot classification tasks in information extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12423–12434, Torino, Italia. ELRA and ICCL.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Tag breakdown for UD-NewsCrawl

Table 5 and Table 6 shows the distribution of Universal Part-of-Speech (UPOS) and Dependency Relations (DEPREL) tags for all splits of UD-NewsCrawl.

	Train	Dev	Test
NOUN	41722	5450	5543
ADP	33003	4207	4292
PROPN	32034	3929	3858
VERB	28361	3802	3668
PART	27899	3673	3790
PUNCT	23384	2950	2978
ADV	21029	2797	2718
DET	19515	2582	2503
PRON	15555	1954	1899
ADJ	9173	1099	1132
CCONJ	6621	891	852
SCONJ	6254	794	762
NUM	6088	756	790
INTJ	175	34	31
SYM	122	32	19
X	56	9	7

Table 5: UPOS (Universal Part-of-Speech) distribution

B Annotation Process

B.1 Annotation Guidelines

We highlight some aspects of the annotation guidelines in this section. The full annotation guidelines will be made available after the review period.

B.1.1 Tokenization and lemmatization

Multiword tokens. Since the basic units of annotation in UD are syntactic words, we systematically split off clitics in case they are written attached to their host. In Tagalog, this only concerns **multiword tokens**, in which the linker *-ng* when it is attached to a vowel-final word.

Foreign words. For English or other foreign words mixed in the data, we do not identify the root words. Exceptions are when they are used with Tagalog affixes:

- *nag-e-enjoy* (enjoying) → enjoy
- *nakipag-meeting* (had a meeting) → meeting
- *kaka-check* (just checked, checking) → check

	Train	Dev	Test
case	45243	5800	5870
punct	23360	2948	2973
det	19801	2632	2568
advmod	19690	2610	2529
nsubj	18322	2389	2281
nmod	16012	1997	2074
mark	15375	2031	2032
flat	14925	1814	1782
obl	12953	1691	1735
root	12495	1561	1563
nmod:poss	7780	1025	1039
advcl	7113	917	840
conj	6865	947	886
cc	6497	860	829
obj:agent	6239	765	735
amod	6131	706	773
acl:relcl	5359	715	743
obj	4879	583	600
discourse	3622	492	501
fixed	3492	487	503
nummod	3191	417	401
ccomp	2982	362	392
xcomp	1722	293	241
compound	1464	286	307
parataxis	1340	165	164
appos	1334	176	187
dep	722	71	81
dislocated	710	84	79
compound:redup	439	44	49
list	322	17	16
acl	317	51	50
vocative	113	13	15
goeswith	79	6	3
orphan	77	2	1
reparandum	10	1	0
iobj	9	0	0
obl:agent	6	0	0
cop	1	0	0
csubj	1	0	0
expl	1	0	0

Table 6: DEPREL (Dependency relations) distribution

On some multiword expressions (MWEs). In Tagalog, some MWEs are now commonly contracted into a single word, such as *kundi* (previously *kung hindi*, “if not”) or *anuman* (previously *ano man*, “whatever”). We have opted not to separate these expressions further, and we instead retain them as individual words with lemmas as they were written in the text.

B.1.2 POS Tags

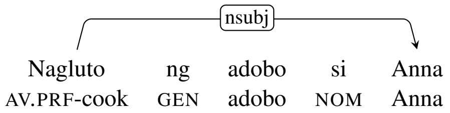
In Tagalog, POS tagging can be tricky and confusing. In general, the POS of words are assigned based on the derived word rather than the root. For example, in *paglakad* (Lemma=*lakad*, "walk"), the root *lakad* may be used as a noun or as a verb. But, if combined with the affix *pag-*, then it can only be used as a noun. For words that appear in their bare forms, we consult the Tagalog-English Dictionary (English, 1986). In general, if a word denotes an object, then they are labelled as a `NOUN`, `ADJ`, or `VERB`.

B.1.3 Morphological Features

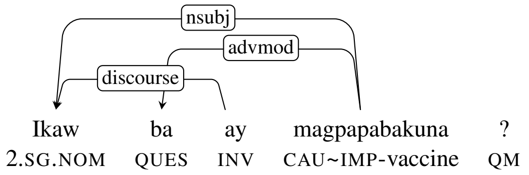
In UD-NewsCrawl, we mark the morphological features that affixes add to the word. Hence, in the sentence *"Lakad na tayo."* (Let's walk), there is no need to mark any features on *lakad* (walk) which is used in its bare root form. In addition, Recent Perfective verb forms do not assign focus, and they are not marked with the `Voice` feature.

B.1.4 Dependency Relations

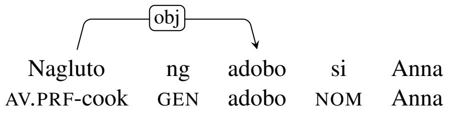
Annotating clausal relations. In UD-NewsCrawl, we annotate the *ang*-marked argument or the nominative as `nsubj`.

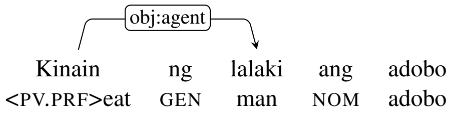


An exception happens if there is an *ay* inversion marker present in the clause. In this case, the first NP is treated as the subject and the second constituent is treated as the predicate.

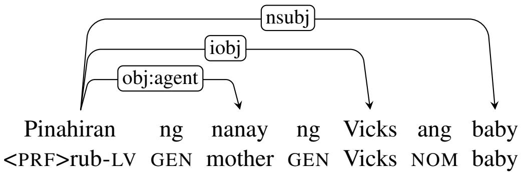


For *ng*-marked clauses, we label patient or undergoer arguments as `obj`, and label non-subject agents as `obj:agent`.

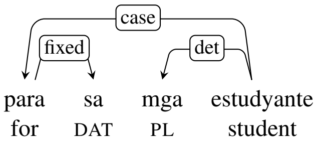




The UD framework only allows one argument to be labelled as the object in a sentence. If there are cases where there are two *ng*-marked arguments, we mark the more agent-like argument as `obj` and the more patient-like argument as `iobj`.



Annotating MWEs. Multiword expressions may be marked as having any of the three relations: `fixed`, `flat`, or `goeswith`. We use `fixed` for fixed grammatical MWEs:



We use `flat` for foreign names, titles, and dates. Finally, we use `goeswith` to link tokens that are typically joined as a single word but are separated due to the original text having typos or errors.

B.2 Differences with UD guidelines

Our annotation guidelines depart from the existing guidelines that were present in UD-TRG and UD-Ugnayan. We highlight these differences below.

POS Tags

- Among the 17 UPOS tags, AUX was not utilized. Some words that express modal meaning (*dapat*, *puwede*) or negation (*huwag*, *hindi*) were marked as ADV instead.
- DET was only used for nominative markers *ang*, *si*, *sina*, the plural marker *mga*, and English articles *the*, *a*, *an*.
- The ADP tag was used for genitive markers *ng*, *ni*, *nina* and oblique/dative markers *sa*, *kay*, *kina* / *kila*.

Morphological Features

- **Nominal and Pronominal Features**
  - `Gender` features were not indicated.
  - The token *sarili* is not analyzed as a pronoun. No reflexive pronouns were identified in the corpus.
- **Modifier Features**

  - We use Degree=Abs for *napaka-*.
  - We use Degree=Equ for *kasin-/kasing*.
  - Polarity=Pos was not used in the corpus.

- **Verb Features**
  - Two Mood values were used in the corpus: Ind (which corresponds with the *realis* mood) and Pot (which corresponds with the *irrealis* mood).
  - Two Aspectual values were used in the corpus: Imp (may or may not have been initiated and not yet complete) and Perf (completed at a certain point of time); habitual and prospective were not used.

- **Other Features**
  - Link feature is not used in the corpus and the *-ng* linker is treated as a separate token (rather than as a suffix).
  - NumType feature is also indicated in the corpus.

Dependency Relations

- nsubj subtypes were not indicated in the corpus.
- csubj was not used when a nominative marker introduces a voice-marked form to reflect the flexibility of syntactic categories in Tagalog.
- case was also used to mark the relation between a linker to a modifier.

C Additional description of components

We train the models using the spaCy framework, resulting in a multi-task pipeline that consists of several components. In this section, we provide additional description of these components.

Lemmatizer We use a recursive edit-tree lemmatizer based on Müller et al. (2015) that derives lemmatization rules from a set of examples. For a reasonably-sized corpus, this algorithm produces thousands of edit-trees for each token-lemma pair. In order to pick the correct edit-tree, we train a small network that learns to predict the most-probably edit-tree to lemmatize the token. The network architecture consists of a convolutional neural network (CNN) and a layer-normalized maxout activation function. We set the minimum frequency of an edit tree to 3, and used the surface form of the token as a backoff when no applicable edit-tree is found.

Morphological and POS tagger We employ a standard statistical approach for both morphological annotation and POS tagging by treating it as a token classification task. A vector of tag probabilities are predicted for each token in the batch, and the most probable tag is selected for each token. The network is then optimized using a categorical cross-entropy loss. For the morphological analyzer, we made every unique combination of morphological features as a class. The main limitation of this approach is that it can only predict feature combinations that exist in the training set.

Dependency parser We employ a transition-based approach to dependency parsing as described in Honnibal and Johnson (2015). We have not explored graph-based algorithms such as the biaffine model from Dozat et al. (2017), and we leave that for future work.

D Architecture and training hyperparameters

D.1 Architectures

The word-embedding models use spaCy's default token-to-vector encoding network that consists of embedding and contextual encoding subnetworks. The hyperparameters are shown in Table 8.

Parameter	Value
Dropout	0.1
Gradient accumulation	1
Patience	1600
Max steps	20000
Optimizer	Adam
β_1	0.9
β_2	0.999
L2	0.01
Gradient clip	1.0
ϵ	1e-7
Learning rate	0.001

Table 7: Training hyperparameters.

For transformer-based models that create context-sensitive vectors, we use the default parameters from HuggingFace and only set the window size for obtaining sequences for transformer processing. For our baseline models, we set a window size of 128 and a stride of 96 characters.

Parameter	Value
Embedding subnetwork	
Row sizes	5000, 1000, 2500, 2500
Attributes	NORM, PREFIX, SUFFIX, SHAPE
Encoding subnetwork	
Width	256
Depth	8
Window size	1
Maxout pieces	3

Table 8: Hyperparameters of the token-to-vector encoding network.

D.2 Training hyperparameters

Table 7 shows the hyperperameters we used for training all baseline models.

D.3 Additional results for baseline models

Additional and fine-grained results for the morphological analyzer and dependency parser can be found in Table 9 and Table 10 respectively.

	No embeddings	fastText	Multi hash emb.	mDeBERTa-v3	RoBERTa (tl)	XLM-RoBERTa
Aspect	83.65	88.73	91.21	91.54	92.83	92.62
Mood	87.87	91.47	92.60	93.43	94.06	93.86
Voice	77.59	79.53	81.38	81.62	83.12	82.08
Case	98.88	98.83	98.99	98.89	98.88	98.95
Number	99.20	99.18	99.33	99.27	99.23	99.34
Person	99.11	99.07	99.31	99.23	99.27	99.42
PronType	97.69	97.79	98.11	98.04	98.21	98.19
NumType	91.72	91.61	91.75	90.75	91.53	91.72
Deixis	94.17	94.17	95.16	94.73	95.43	95.25
Abbr	17.45	21.05	2.90	19.87	9.79	21.05
Polarity	97.52	97.14	98.02	97.90	97.01	98.15
Typo	30.00	32.79	21.82	44.44	52.94	49.32
Degree	65.45	65.38	79.31	78.69	83.87	90.62
Clusivity	98.94	99.47	99.47	99.47	98.67	99.73
PartType	100.00	98.59	92.54	94.12	90.91	85.71
Polite	87.88	97.22	100.00	98.63	100.00	97.22

Table 9: Fine-grained morphological analysis F1-score results for all baseline models.

	No embeddings	fastText	Multi hash emb.	mDeBERTa-v3	RoBERTa (tl)	XLM-RoBERTa
root	83.66	84.77	85.53	86.60	86.65	88.09
advmod	78.87	79.37	81.24	81.67	82.85	82.52
case	89.29	90.07	90.56	90.31	90.84	90.87
compound	27.62	30.44	22.61	39.41	38.42	33.68
obj	74.36	74.04	76.43	79.80	82.64	81.41
det	93.63	93.23	93.76	94.09	94.12	94.60
nsubj	80.80	81.50	82.31	83.57	85.75	87.07
nmod:poss	76.39	78.55	79.27	79.10	82.19	81.84
obl	65.50	67.71	68.64	71.21	73.55	74.65
cc	83.84	84.68	83.78	87.52	88.70	88.30
conj	61.16	61.99	65.87	71.51	71.44	76.03
fixed	72.71	69.80	72.77	75.23	74.80	75.79
amod	72.30	73.22	77.43	75.51	78.09	77.89
mark	78.25	79.54	79.37	82.37	83.84	84.13
acl:relcl	64.07	64.11	66.49	69.01	71.59	75.18
nmod	55.03	56.96	57.82	60.20	63.73	63.72
ccomp	55.39	56.65	57.04	60.14	67.91	69.22
advcl	49.08	50.45	49.72	57.14	61.22	60.88
flat	82.27	82.36	83.38	84.48	84.01	86.09
discourse	76.34	74.62	76.28	79.15	82.01	84.01
parataxis	29.69	31.11	30.02	37.74	36.88	40.79
nummod	80.33	83.40	83.94	86.11	85.10	85.61
obj:agent	78.46	81.90	83.24	84.39	87.90	85.90
dep	9.52	14.01	23.81	27.23	29.83	22.32
compound:redup	47.83	35.79	43.90	56.60	56.82	73.68
xcomp	54.23	52.86	58.17	58.35	65.57	61.47
appos	53.03	44.44	58.29	64.69	60.85	63.61
acl	6.78	9.38	17.86	10.67	10.00	21.33
list	12.66	16.67	17.39	10.17	17.54	17.24
vocative	36.36	58.33	16.22	71.43	62.07	81.48
dislocated	19.18	28.37	31.58	26.49	28.36	23.53
goeswith	0.00	0.00	0.00	0.00	0.00	0.00
orphan	0.00	0.00	0.00	0.00	0.00	0.00

Table 10: Fine-grained dependency parsing per-type LAS results for all baseline models.

E Cross-lingual performance on UD-NewsCrawl

Set-up In order to understand how dependency parsers trained in another language perform on the test split of UD-NewsCrawl, we follow the setup described in Aquino and de Leon (2020) and evaluate parsers trained on languages that have a sufficiently large treebank (≥50k words) and are linguistically similar to Tagalog based on a metric defined by Agić (2017) using features from the World Atlas of Language Structures (WALS, Dryer and Haspelmath, 2013).

Given a source language S and target language T, we obtain the WALS measure $\text{dist}_W(S,T)$ by calculating the Hamming distance d_H of the WALS feature vectors $\mathbf{v}_S$ and $\mathbf{v}_T$, normalized with respect to the number of features $f_{S,T}$:

$$\text{dist}_W(S,T) = \frac{d_h(\mathbf{v}_S, \mathbf{v}_T)}{f_{S,T}}$$

For Tagalog, the most linguistically similar languages with available treebanks are Indonesian (UD-GSD), Vietnamese (UD-VTB), Romanian

(UD-RRT), Catalan (UD-AnCora), and Ukrainian (UD-IU). Table 11 shows the performance of a parser trained on each language and evaluated on the test split of UD-NEWSCRAWL.

Results We find that there is limited cross-lingual generalization, with models trained on other languages achieving significantly lower performance on Tagalog parsing. This suggests that Tagalog's unique syntactic patterns and typological features may require dedicated resources for effective parsing, rather than relying on transfer from other languages. While previous studies on smaller Tagalog treebanks show more promising cross-lingual transfer (Aquino and de Leon, 2020; Miranda, 2023b), our results on the substantially larger UD-NEWSCRAWL may provide a more reliable assessment of the limitations of cross-lingual generalization.

F Examples for each topic in UD-NEWSCRAWL

In this section, we illustrate some examples from UD-NEWSCRAWL that correspond to different topics as defined by SIB-200 (Adelani et al., 2024).

- **Entertainment**
  - *Alam din daw niyang masaya na si Nikki sa boyfriend nitong si Billy Crawford.* (He said he also knows that Nikki is already happy with her boyfriend Billy Crawford.)
  - *Out of the blue, sumali kami sa kaswal na chikahan nina Allan at direk Dom.* (Out of the blue, we joined in the casual chat between Allan and director Dom.)
  - *Sa issues na tatalakayin sa show, aamin na kaya si Bianca King kung siya na ang bagong Reyna ng Kasungitan na ibinabato sa kanya?* (Among the issues that will be discussed on the show, will Bianca King finally admit if she is the new Queen of Anger that is being thrown at her?)
- **Health**
  - *Kailangan umanong masuportahan ito ng dokumento o ebidensya, tulad ng resulta ng drug test.* (This must be supported by documents or evidence, such as drug test results.)
  - *Kasi tumaba na ako, ang hirap kayang magpapayat.* (Because I've gained weight, it's hard to lose weight.)
  - *Noong 1983 pa nagkaroon na ng product recall ang Saridon matapos mapatunayan ng US FDA na maaari itong pagmulan ng cancer sa mga taong umiinom ng mga gamot na may phenacetin (kasama na ang Saridon).* (Saridon had a product recall as early as 1983 after the US FDA confirmed that it could cause cancer in people taking medications containing phenacetin (including Saridon).)
- **Sports**
  - *Ititiklop ng Elasto Painters ang eliminations kontra Alaska bukas sa Ynares Center sa Antipolo.* (The Elasto Painters will repeat the eliminations against Alaska tomorrow at the Ynares Center in Antipolo.)
  - *Simula na rin ng banggaan ng defending NBA champion at Eastern Conference No. 2 seed Boston laban sa 7th seeded Chicago Bulls.* (The clash between defending NBA champion and Eastern Conference No. 2 seed Boston and the 7th seeded Chicago Bulls has begun.)
  - *Dalawang larong liban si Taulava sa pagsaklolo sa Smart Gilas Pilipinas na pumang-anim nga lang sa katatapos na Guangzhou Asian Games.* (Taulava missed two games to help Smart Gilas Pilipinas, which only finished sixth in the recently concluded Guangzhou Asian Games.)
- **Politics**
  - *Ang tanging kalaban ni Arroyo sa nasabing posisyon ay ang private citizen na si Adonis Simpao.* (Arroyo's only opponent in the said position is private citizen Adonis Simpao.)
  - *Samantala, mabilis na itinanggi ni Sen. Jinggoy ang isyung ang kampo nila ang nagpakalat na bayad ang pag-eendorso ni Dolphy kay presidential aspirant Manny Villar.* (Meanwhile, Sen. Jinggoy quickly denied the issue that his camp spread the rumor that Dolphy's endorsement of presidential aspirant Manny Villar was paid for.)

			Lemm.	**UPOS**	**Morph.**		**Dep. Parsing**	
Language	**Treebank**	dist$_W$	*Acc*	*Acc*	*Acc*	*F1-score*	*UAS*	*LAS*
Indonesian	UD-GSD	0.446	65.6	40.3	50.1	7.4	26.2	7.6
Ukrainian	UD-IU	0.455	73.8	11.3	8.8	2.7	14.2	2.2
Vietnamese	UD-VTB	0.469	62.8	30.6	12.2	4.3	22.8	7.1
Romanian	UD-RRT	0.471	64.6	34.1	21.8	4.1	27.9	9.2
Catalan	UD-AnCora	0.472	71.2	33.2	21.3	5.7	25.5	6.8

Table 11: Performance of a pipeline trained on a treebank of the top five languages typologically similar to Tagalog using a WALS-based metric dist$_W$ from Agić (2017), as evaluated on the test split of UD-NewsCrawl.

  - *Ang kapatid ngayon ni Nograles na si Jose ang presidente ng PDIC.* (Nograles' brother Jose is now the president of PDIC.)

- **Travel**
  - *Dumating kami dito sa Gold Coast noong Martes ng umaga mula sa mahabang biyahe mula sa Maynila.* (We arrived here on the Gold Coast on Tuesday morning after a long trip from Manila.)
  - *Ang sabi sa amin, ang biyahe mula Maynila hanggang Santa Ana Park ay inaasahang iiksi ng isang oras na lamang kapag nagawa at napadaanan na ang nasabing highway.* (We are told that the trip from Manila to Santa Ana Park is expected to be shortened to just one hour once the said highway is completed and passable.)
  - *Hindi gaanong malalaki ang mga gusali at limitado hanggang alas-siyete ng gabi ang operasyon ng mall nila doon.* (The buildings are not very large and their mall's operations are limited to seven o'clock in the evening.)

- **Geography**
  - *Malaysia, Singapore, at ibang karatig-bansa.* (Malaysia, Singapore, and other neighboring countries.)
  - *Ang mga puno sa magkabilang hanay ng MacArthur Highway mula City of San Fernando hanggang sa Angeles City ay nagbibigay ng sariwa at luntiang damdamin sa sinumang dumadaan dito.* (The trees on both sides of the MacArthur Highway from the City of San Fernando to Angeles City give a fresh and green feeling to anyone passing by.)
  - *Agrikultura ang major industry ng lugar kaya't milya-milya ng lupaing tinamnan sa palay, maize, niyog at marami pang iba ang makikita ng bisitang nagbabaybay ng probinsya.* (Agriculture is the area's major industry, so visitors traveling through the province will see miles and miles of land planted with rice, maize, coconuts, and many other crops.)

- **Science/Technology**
  - *Ginawan ng maraming pag-aaral tulad ng pagsasagawa ng Parañaque Spillway na bibigyang-daan ang daloy ng tubig galing sa Laguna Lake papuntang Manila Bay.* (Many studies have been conducted, such as the construction of the Parañaque Spillway, which will allow the flow of water from Laguna Lake to Manila Bay.)
  - *Ang PRES ay isang electric service na gagamit ng prepaid metering system na layong payagan ang mga residential customers na bumili ng credit o load para sa paggamit ng kuryente hanggang sa ito'y maubos.* (PRES is an electric service that will use a prepaid metering system that aims to allow residential customers to purchase credit or load for electricity use until it is exhausted.)
  - *Ayon kay Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA) Director Prisco Nilo, karaniwang nagaganap ang meteor shower tuwing buwan ng Agosto.* (According to Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA) Director Prisco Nilo, meteor showers usually occur every August.)

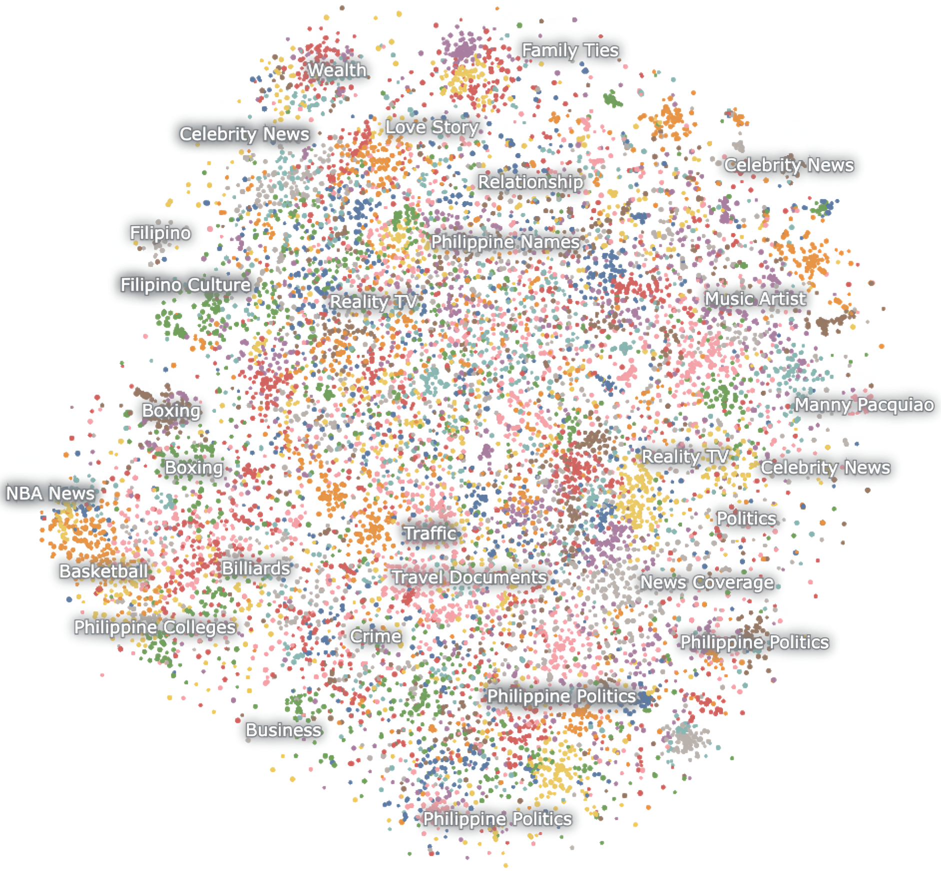


Figure 4: Embedding map generated using the Nomic Atlas API for fine-grained topic classification.

G Fine-grained topic classification

We also perform fine-grained topic classification using a multilingual generate text embedding (GTE) model (Zhang et al., 2024). We then use the Nomic Atlas API[4] to perform topic modelling using a hierarchical clustering model. The embedding map can be found in Figure 4.

The embedding map suggests that UD-NEWSCRAWL captures multiple domains of Philippine society, with 20 major topic clusters identified through our analysis. These clusters encompass several key categories: entertainment (Celebrity News, Reality TV, Love Story), sports (Basketball, NBA News, Boxing), politics (Philippine Politics appearing in multiple distinct clusters), and cultural elements (Filipino Culture, Philippine Names).

Our findings suggest that UD-NEWSCRAWL captures a diverse range of topics in Philippine news media, with particularly strong representation of entertainment, sports, and political content. This topical distribution, as revealed through the embeddings, is consistent with our earlier topic analysis findings in §6.3, reflecting the typical content priorities and coverage patterns of mainstream Philippine news outlets.

H Prompt template for topic classification

We provide the prompt template for the topic classification experiment in Figure 5. We find that writing the instructions in English (instead of Tagalog) and specifying the language of the text-in-question ("The text is in Tagalog...") gives more consistent and parseable outputs in our evaluation.

I Error analysis

During quality control (§4.4), we compare our expert annotations with the existing global and language-specific rules in the UD framework using the UD validator. We find that there are two major error categories in our expert annotations: incompatibilities between particular UPOS tags (L3) and language-specific labels that do not yet exist in the UD guidelines (L4).[5]

[4] https://atlas.nomic.ai/

[5] https://universaldependencies.org/validation-rules.html

PROMPT TEMPLATE
Given the text below, classify it into one of the following topics:

science/technology
travel
politics
sports
health
entertainment
geography

The text is in Tagalog, but your response should be in English. Return only the topic name in small caps.

Below are some examples:

Text: Sa simula ng digmaan kalimitang naglakbay ang mga ito sa ibabaw ng dagat, nguni't noong nagsimula nang sumulong ang radar at nagiging mas tumpak na ito, ang mga submarino ay napilitang sumisid sa ilalim ng tubig upang maiwasang may makakita sa mga ito.
Answer: science/technology

Text: Noong 1994, humantong ang di-pagkakasundo sa paglikha ng nag-aangking Republika ng Transnitria sa silangang Moldova, na may sariling pamahalaan at pera subalit hindi kinikilala ng anumang miyembrong bansa ng UN.
Answer: politics

Here is the text you need to classify:

Text: {text}
Answer:

Figure 5: Prompt template for topic classification.

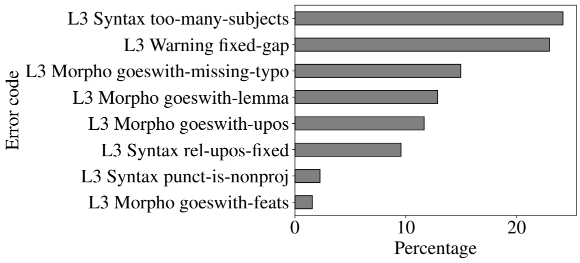


Figure 6: Breakdown of L3 errors that relate to incompatibilities between particular UPOS tags as detected by the official UD validator.

I.1 Type L3 Errors

Figure 6 show a breakdown of L3 errors. We find that some of these are due to the annotation decisions we made regarding Tagalog grammar as described in §3.2 (and Appendix B) which contradict some aspects of the universal guidelines.

I.2 Type L4 Errors

The majority of L4 errors include language-specific labels that affect morphological features and dependency relations. For example, the validator might say that a morphological feature is not permitted with a certain UPOS (e.g., "Feature Case is not permitted with UPOS DET in language [tl]"). Resolving these errors involve either one or two fixes: update the language-specific guidelines to reflect new cases found in our treebank or correct the expert annotations. We find that most of the fixes involve the former approach, as the language-specific guidelines for Tagalog are still sparse.

This pattern is entirely expected, given that the existing Tagalog treebanks are relatively small in size and scope. The current language-specific guidelines were likely derived from these limited samples, which may not capture the full range of linguistic phenomena present in Tagalog. Our larger and more diverse treebank effectively serves as a window into previously undocumented morphosyntactic patterns, revealing cases where determiners, for instance, can indeed carry case marking in specific contexts. This presents a valuable opportunity to expand and refine the language-specific guidelines, making them more comprehensive and representative of actual language usage. Rather than viewing these L4 errors as mere validation issues, they can be seen as important signals highlighting areas where our understanding of Tagalog's formal grammatical structures needs to be updated and codified in the Universal Dependencies framework.