Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Lexicon-Guided Morphological Tag Injection for Low-Resource Filipino-Cebuano Neural Machine Translation

Kristine Mae M. Adlaon[†,*], Nelson Marcos[‡]
[†] University of the Immaculate Conception, Davao City, Philippines
[‡] De La Salle University, Manila City, Philippines

Abstract

Neural Machine Translation (NMT) remains difficult for low-resource languages, especially those with complex word formation systems. This work focuses on the Filipino–Cebuano language pair, where verbs encode voice and aspect using different morphological patterns. Although the two languages are closely related, their distinct verb formation strategies often create ambiguity and mismatches during translation, leading to errors in predicate interpretation and grammatical alignment. Pretrained multilingual models such as NLLB-200 provide broad language coverage, but they frequently struggle with predicate-level accuracy in closely related Philippine languages due to insufficient explicit morphological grounding. We propose a lexicon-guided morphological tag injection framework that enriches source-side input with structured linguistic cues, including aspect and voice markers derived from a curated morphological lexicon. Rather than modifying the model architecture or introducing new token embeddings, we inject morphological metadata directly into the input sequence and perform parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). Experimental results show consistent improvements over baseline fine-tuning, particularly in constructions involving complex verbal morphology and one-to-many or many-to-one lexical mappings.

Keywords: Low-Resource Neural Machine Translation, Morphologically Rich Languages, Lexicon-Guided Tag Injection, Low-Rank Adaptation (LoRA), Multilingual Transformer Models, Filipino–Cebuano Translation

1. Introduction

Neural machine translation (NMT) has achieved substantial performance gains in recent years through the adoption of transformer-based architectures and large-scale multilingual pretraining. Models such as MarianNMT [1], mBART [2], mT5 [3], M2M-100 [4], and Meta's No Language Left Behind (NLLB) [5] have demonstrated strong cross-lingual transfer capabilities, particularly when trained on massive multilingual corpora spanning hundreds of languages [2, 5–7]. Despite these advances, translation quality for low-resource and morphologically rich languages remains uneven, especially for language pairs that are underrepresented in training data.

Filipino (a language heavily based on Tagalog) and Cebuano are two major Philippine languages with distinct but related morphological systems [8–10]. Tagalog exhibits rich verbal morphology characterized by extensive affixation and a voice or focus system that encodes grammatical relations directly on the verb [8, 9], while Cebuano relies on a different set of morphological cues and syntactic constructions, often expressed through verbal prefixes, particles, and clause-level structure rather than extensive infixation [10, 11]. Although both languages are spoken by millions, parallel corpora for Tagalog-Cebuano remain limited [12–14], and existing multilingual machine translation models often struggle to capture fine-grained morphological distinctions that are critical for accurate translation in low-resource settings [5, 15, 16]. As a result, translations may suffer from errors in tense, aspect, voice, and argument structure, phenomena closely tied to morphology [17].

In this work, we propose a morphology-aware machine translation approach for Filipino–Cebuano that combines lexicon-guided morphological tagging with parameter-efficient

* kadlaon@uic.edu.ph

fine-tuning of a task-specific machine translation pretrained language model. Instead of introducing new tokens or modifying the model architecture, we inject morphological tags derived from a curated Filipino(Tagalog) lexicon directly into the source text. This strategy preserves compatibility with the pretrained tokenizer while exposing the model to explicit morphological cues during training and inference.

2. Morphological Characteristics of Filipino and Cebuano

Filipino (largely based on Tagalog) and Cebuano belong to the Austronesian language family and share several typological characteristics, including rich affixation, focus or voice systems, and relatively flexible word order [8, 10, 18, 19]. Despite these similarities, the two languages differ substantially in how grammatical relations, tense, aspect, and semantic roles are morphologically encoded [8–10]. These differences pose challenges for neural machine translation, particularly in low-resource settings where parallel training data is limited [15, 17].

2.1. Morphology of Filipino (Tagalog)

Filipino exhibits a highly productive derivational and inflectional morphology, most prominently realized through the extensive use of prefixes, infixes, suffixes, and circumfixes [8]. Verbal morphology encodes multiple grammatical dimensions simultaneously, including voice (focus), aspect, and mood, often within a single verbal form [9]. Common affixes such as *mag-*, *-um-*, *-in-*, *i-*, *-an*, and *-hin* interact with reduplication to signal distinctions such as completed versus incompleted aspect, actor versus patient focus, and causative or locative constructions [8, 9].

Many of these morphological markers are not transparently aligned with equivalent surface forms in Cebuano or other Philippine languages, as changes in verbal affixation can alter the syntactic prominence of arguments without a corresponding change in word order [8, 10]. In data-driven neural machine translation systems, particularly those relying on subword segmentation alone, such distinctions may be underrepresented or inconsistently learned, leading to errors in argument structure, voice, or semantic role interpretation [15, 17]. Additionally, Filipino morphology is lexeme-sensitive, where affix selection and attachment may depend on lexical properties of the root [8]. This motivates the use of lexicon-based morphological information, which can provide explicit linguistic cues that are difficult to infer from surface forms alone in low-resource settings [17].

2.2. Morphology of Cebuano

Cebuano also employs affixation and reduplication but differs from Filipino in both the distribution and functional load of its morphological markers [10, 11]. Although Cebuano verbs encode aspect and voice, the system is generally less affix-dense than Filipino, and several grammatical distinctions are expressed through particles or syntactic constructions rather than through rich verbal morphology [10]. Moreover, Cebuano verbal affixes do not map one-to-one with Filipino affixes, even when the underlying semantic roles are similar, resulting in non-isomorphic morphological correspondences between the two languages [8, 10]. As a result, direct translation from Filipino to Cebuano requires not only lexical substitution but also morphological reinterpretation, particularly for verbs and predicate-centered constructions. When such reinterpretation is learned implicitly from limited parallel data, translation quality may degrade, especially for rare or morphologically complex forms [15, 17].

3. Methodology

This study adopts a two-stage experimental methodology. In the first stage, we conducted a comparative baseline experiment across multiple pretrained machine translation models to identify a suitable backbone for morphology-aware experimentation. In the second stage, we implemented and evaluated a lexicon-guided morphology-aware input augmentation approach using the selected model. The architectural pipeline for the morphology-aware translation system shown in Figure 1 begins with the ingestion of a raw Filipino-Cebuano parallel corpus, along with specialized lemma and inflectional lexicons. These resources are processed by a modular *Morphological Tagger and Cleaner*, which performs lexicon-guided tag injection to output a data stream: the augmented source (`src_tagged`) and the corresponding ground-truth target (`tgt`). This enriched input is fed into the `NLLB-200` backbone (`facebook/nllb-200-3.3B`), where the primary model weights are kept frozen to leverage robust pretrained cross-lingual representations. Task-specific adaptation is achieved via *LoRA Adapters* integrated into the attention mechanism, specifically targeting the projection layers q_{proj} and v_{proj}. The training phase utilizes a `Seq2Seq` objective, optimized with 16-bit floating-point precision (`FP16`) and gradient accumulation to maintain computational efficiency while generating the final adapted LoRA weights.

3.1. Preliminary Model Selection Experiment

Before introducing morphological tagging, we conducted a preliminary experiment to assess the performance of several widely used neural machine translation models on the Filipino–Cebuano language pair. The objective of this experiment was to identify a model that provides a sufficiently strong baseline and stable cross-lingual representations to support subsequent morphology-aware input augmentation. Specifically, we evaluated three models: MarianMT, representing task-specific neural machine translation systems trained on curated parallel corpora; M2M-100, a multilingual many-to-many model trained on large-scale parallel data; and NLLB-200, a massively multilingual model explicitly designed to support low-resource languages. All models were evaluated in both translation directions (Filipino $\rightarrow$ Cebuano and Cebuano $\rightarrow$ Filipino) using the same test set and evaluation metrics: BLEU, ChrF++, and COMET. The results show that NLLB-200 substantially outperforms MarianMT and M2M-100 across all evaluation metrics and both translation directions. Notably, the results are consistent across surface-based (BLEU, ChrF++) and semantic (COMET) metrics as shown in Table 1. These findings indicate that NLLB-200 provides a stronger representational foundation for Filipino–Cebuano translation and is therefore better suited for investigating the impact of morphology-aware input augmentation. Based on this preliminary analysis, NLLB-200 was selected as the base model for all subsequent morphology-aware experiments.

3.2. Model Fine-Tuning

The resulting morphology-aware translation system, which we refer to as *iWag*, was built upon the selected NLLB-200 backbone. The NLLB-200 model was fine-tuned on the tagged parallel corpus using a parameter-efficient fine-tuning strategy based on Low-Rank Adaptation (LoRA). LoRA adapters were applied to the attention projection layers of the transformer, enabling task-specific adaptation while keeping the majority of model parameters frozen. This setup substantially reduces computational cost and mitigates overfitting risks, which are particularly relevant for low-resource language pairs. Training was conducted using the Hugging Face Transformers framework, with mixed-precision training enabled where hardware permitted. Hyperparameters were selected to balance training stability and computational efficiency.

3.3. **Evaluation**

To know the effect of morphological tagging, the fine-tuned model was evaluated under two input conditions. In the tagged condition, lexicon-guided morphological tags were injected into the source sentences, while in the untagged condition, the same sentences were provided in their original, unmodified form. Both conditions used the identical trained model and decoding configuration to ensure a controlled comparison. Translation quality was assessed using BLEU, ChrF++, and COMET, capturing complementary surface-level and semantic aspects of translation performance. Performance differences between the tagged and untagged inputs were analyzed to quantify the impact of morphology-aware input augmentation. To further contextualize the quantitative evaluation, we conducted a controlled verb-focused analysis to examine how morphology-aware input augmentation affects the translation of aspect and voice across Filipino–Cebuano and Cebuano–Filipino directions. The experiment consisted of manually evaluating 300 sentences per aspectual category (complete, progressive, and contemplative) and per translation direction, with outputs categorized as correct, partially correct, or incorrect with respect to verb focus and aspect realization.

4. **Results and Discussion**

4.1. **Results**

Table 2 and Table 3 reports the results of a verb-focused manual evaluation comparing iWag against two strong baselines (GPT-5 and GNMT) across Filipino–Cebuano and Cebuano–Filipino translation directions. For each direction, 300 sentences were evaluated per aspectual category (complete, progressive, and contemplative) and classified as correct, partially correct, or incorrect with respect to verb focus and aspect realization.

In the Filipino–Cebuano direction, iWag consistently outperformed both baselines across all aspectual categories. For complete aspect constructions, iWag achieved a correctness rate of 90%, substantially higher than GPT-5 (66%) and GNMT (71%). This trend persisted for contemplative forms, where iWag reached 87% correctness, again exceeding both comparison systems by a large margin. The advantage of morphology-aware augmentation was most pronounced for progressive constructions, which are known to involve complex interactions between affixation and reduplication in Filipino. While iWag correctly translated 70% of progressive cases, GPT-5 and GNMT struggled considerably, with correctness rates of only 26% and 15%, respectively. These results suggest that explicit morphological cues enable the model to better preserve aspectual meaning during translation, particularly for forms that are difficult to infer from surface patterns alone.

The reverse translation direction proved more challenging overall, reflecting the greater morphological complexity of Filipino verbal forms. Nonetheless, iWag remained competitive and, in several cases, superior to the baselines. For complete aspect constructions, iWag achieved a 60% correctness rate, outperforming GNMT (51%) and slightly exceeding GPT-5 (57%). For progressive and contemplative forms, performance gaps between systems narrowed, with GPT-5 marginally outperforming iWag in correctness; however, GPT-5 also exhibited a substantially higher proportion of partially correct outputs. This pattern suggests that while large general-purpose models may approximate the intended meaning, they often fail to produce fully well-formed morphological realizations. In contrast, iWag demonstrated more stable control over verb morphology, as reflected in lower rates of severe errors for several categories.

Across both translation directions, progressive aspect emerged as the most error-prone category for all systems as shown in Table 4. This aligns with linguistic expectations, as progressive constructions in Filipino frequently involve reduplication and infixation, which are

structurally more complex than perfective forms and therefore more difficult for subword-based models to capture without explicit morphological guidance. A recurring error pattern observed in the outputs—particularly in the Cebuano direction—is the neutralization of imperfective or progressive aspect into perfective forms. In several cases, the system preserved lexical meaning but failed to realize the appropriate progressive morphology, instead generating a completed form.

A complementary pattern was observed in the Cebuano-to-Filipino direction, where contemplative or future-oriented constructions were frequently reduced to completed forms as shown in Table 5. In several cases, Cebuano verbs marked with contemplative morphology (e.g., *mo-*, *maka-*) were translated into Filipino perfective forms rather than future or contemplative equivalents. Although lexical meaning was often preserved, the temporal or modal interpretation shifted from intended futurity to completed action.

The consistently lower error rates observed for iWag indicate that lexicon-guided morphological tagging helps disambiguate these constructions by providing the model with direct access to aspectual and voice-related information. Moreover, the reduction in incorrect outputs—especially in the TL→CEB direction—demonstrates that morphology-aware input augmentation improves not only lexical choice but also the faithful realization of grammatical meaning.

5. Conclusion

These results show that explicit morphological augmentation yields substantial gains in translating aspect- and voice-sensitive constructions, particularly in the low-resource Filipino–Cebuano setting. While large pretrained models exhibit strong generalization capabilities, they remain prone to systematic errors when faced with non-isomorphic morphological mappings. By contrast, iWag's lexicon-guided approach enables more consistent morphological reinterpretation, leading to higher accuracy and fewer severe errors. These findings complement the automatic metric results and underscore the importance of incorporating lightweight linguistic knowledge into modern neural machine translation systems.

Acknowledgements

The researchers express their gratitude to the Department of Science and Technology - Philippine Council for Industry, Energy and Emerging Technology Research and Development (DOSTPCIEERD) for providing the funding to purchase the equipment used in training the models.

References

[1] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch. "Marian: Fast Neural Machine Translation in C++". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*. Association for Computational Linguistics, 2018.

[2] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. "Multilingual Denoising Pre-training for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 726–742.

[3] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)* (2021).

[4] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, M. Elbayad, Y. Wang, S. Edunov, E. Grave, A. Joulin, and M. Auli. "Beyond English-Centric Multilingual Machine Translation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2021.

[5] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, Y. He, et al. "No Language Left Behind: Scaling Human-Centered Machine Translation". In: *arXiv preprint arXiv:2207.04672* (2022).

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[7] R. I. Baliber, C. Cheng, V. Mamonong, and K. M. M. Adlaon. "Bridging Philippine Languages With Multilingual Neural Machine Translation". In: *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*. Ed. by A. Karakanta, A. K. Ojha, C.-H. Liu, J. Abbott, J. Ortega, J. Washington, N. Oco, S. M. Lakew, T. A. Pirinen, V. Malykh, V. Logacheva, and X. Zhao. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 14–22. URL: https://aclanthology.org/2020.loresmt-1.2.

[8] P. Schachter and F. T. Otanes. *Tagalog Reference Grammar*. University of California Press, 1972.

[9] P. R. Kroeger. *Phrase Structure and Grammatical Relations in Tagalog*. Center for the Study of Language and Information, 1993.

[10] M. Tanangkingsing. *A Functional Reference Grammar of Cebuano*. De La Salle University Press, 2009.

[11] C. Rubino. *Ilocano and Cebuano*. Ed. by N. P. Himmelmann and K. A. Adelaar. Routledge, 2000.

[12] K. M. M. Adlaon and N. Marcos. "Neural Machine Translation for Cebuano to Tagalog with Subword Unit Translation". In: *2018 International Conference on Asian Language Processing (IALP)*. 2018, pp. 328–333. DOI: 10.1109/IALP.2018.8629153.

[13] J. L. Fernandez and K. M. M. Adlaon. "Exploring Word Alignment towards an Efficient Sentence Aligner for Filipino and Cebuano Languages". In: *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*. Ed. by A. K. Ojha, C.-H. Liu, E. Vylomova, J. Abbott, J. Washington, N. Oco, T. A. Pirinen, V. Malykh, V. Logacheva, and X. Zhao. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 99–106. URL: https://aclanthology.org/2022.loresmt-1.13.

[14] K. M. M. Adlaon and N. Marcos. "Building the Language Resource for a Cebuano-Filipino Neural Machine Translation System". In: *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*. NLPIR '19. Tokushima, Japan: Association for Computing Machinery, 2019, 127–132. ISBN: 9781450362795. DOI: 10.1145/3342827.3342833. URL: https://doi.org/10.1145/3342827.3342833.

[15] P. Koehn and R. Knowles. *Six Challenges for Neural Machine Translation*. 2017.

[16] K. M. M. Adlaon and N. Marcos. "Finding the Optimal Byte-Pair Encoding Merge Operations for Neural Machine Translation in a Low-Resource Setting". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 14673–14682. DOI: 10.18653/v1/2024.findings-emnlp.860. URL: https://aclanthology.org/2024.findings-emnlp.860/.

[17] R. Sennrich and B. Haddow. "Linguistic Input Features Improve Neural Machine Translation". In: (2016).

[18] K. A. Adelaar and N. P. Himmelmann. *The Austronesian Languages of Asia and Madagascar*. Routledge, 2005.

[19] C. Cheng, K. M. M. Adlaon, M. Aquino, E. Fernandez, and K. Villanueva. "MAG-Tagalog: A rule-based Tagalog morphological analyzer and generator". In: *n Proceedings of the 17th philippine computing* (2017), pp. 171–178.

Appendix A. Figures and Tables

A.1. Morph-Aware Translation Pipeline

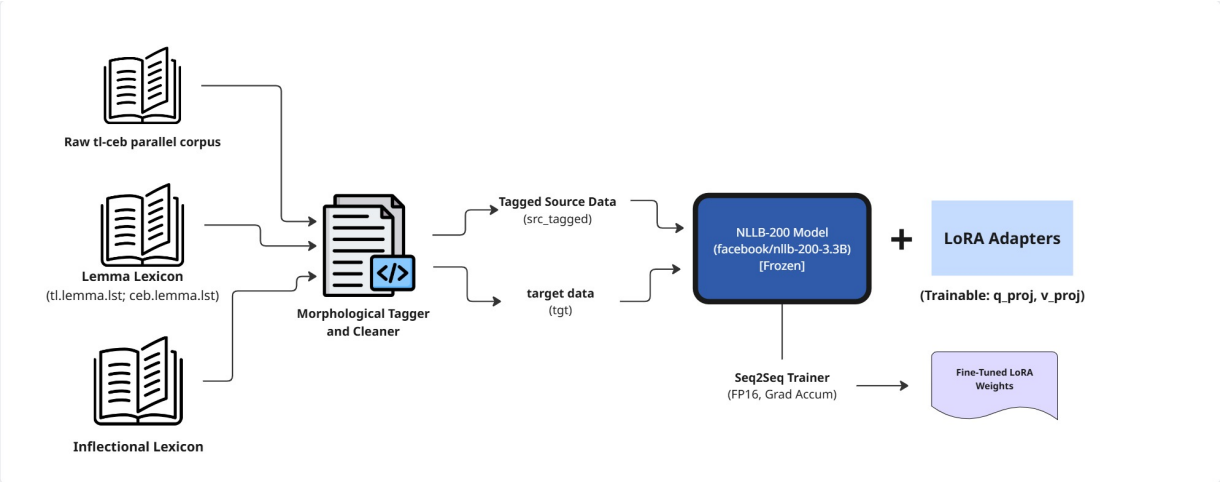

Figure 1. Morph-Aware Translation Pipeline

A.2. Baseline Model Comparison

Table 1. Baseline translation performance across models

Model	Direction	BLEU	ChrF++	COMET
MarianMT	tl → ceb	15.99	44.01	0.6574
	ceb → tl	17.71	45.05	0.6990
M2M-100	tl → ceb	12.20	33.22	0.5498
	ceb → tl	14.99	37.83	0.5451
NLLB-200	tl → ceb	**27.45**	**52.44**	**0.7022**
	ceb → tl	**27.99**	**53.50**	**0.7679**

A.3. Verb Focus Evaluation Results for Filipino→Cebuano Translation

Table 2. Verb Focus Evaluation Results for Filipino→Cebuano Translation

Aspect	Outcome	iWag		GPT-5		GNMT	
		Count	Prop.	Count	Prop.	Count	Prop.
Complete	Correct	270	0.90	198	0.66	214	0.71
	Partially Correct	7	0.02	0	0.00	0	0.00
	Incorrect	23	0.08	102	0.34	86	0.29
Progressive	Correct	210	0.70	78	0.26	44	0.15
	Partially Correct	16	0.05	1	0.00	0	0.00
	Incorrect	74	0.25	221	0.74	256	0.85
Contemplative	Correct	261	0.87	204	0.68	189	0.63
	Partially Correct	30	0.10	1	0.00	2	0.01
	Incorrect	9	0.03	95	0.32	109	0.36

A.4. **Verb Focus Evaluation Results for Cebuano→Filipino Translation**

Table 3. Verb Focus Evaluation Results for Cebuano→Filipino Translation

Aspect	Outcome	iWag		GPT-5		GNMT	
		Count	Prop.	Count	Prop.	Count	Prop.
Complete	Correct	181	0.60	171	0.57	153	0.51
	Partially Correct	34	0.11	80	0.27	35	0.12
	Incorrect	85	0.28	49	0.16	112	0.37
Progressive	Correct	147	0.49	160	0.53	101	0.34
	Partially Correct	34	0.11	97	0.32	19	0.06
	Incorrect	119	0.40	43	0.14	180	0.60
Contemplative	Correct	130	0.49	164	0.55	112	0.37
	Partially Correct	32	0.11	86	0.29	19	0.06
	Incorrect	138	0.40	50	0.17	169	0.56

A.5. **Examples of Progressive-to-Perfective Neutralization Errors (Tl→Ceb)**

Table 4. Examples of Progressive-to-Perfective Neutralization Errors (Tl→Ceb)

Filipino (Source)	Cebuano (Progressive)	iWag Output
Umiinom siya ng tubig.	Nagainom siya ug tubig.	Nag-inom siya ug tubig.
Naglalakad siya sa tubig.	Nagalakaw siya sa tubig.	Naglakaw siya sa tubig.
Iniinom niya ang tubig.	Ginainom niya ang tubig.	Giinom niya ang tubig.

A.6. **Examples of Contemplative-to-Perfective Neutralization Errors (Ceb→Tl)**

Table 5. Examples of Contemplative-to-Perfective Neutralization Errors (Ceb→Tl)

Cebuano (Source)	Filipino (Contemplative)	iWag Output
Moinom siya ug tubig.	Iinom siya ng tubig.	Uminom siya ng tubig.
Makamugna og nagkalain-laing butang pinaagi sa gawasnon nga pagdibuho.	Makakalikha ng iba't ibang bagay sa pamamagitan ng malayang pagguhit.	Lumikha ng iba't ibang bagay sa pamamagitan ng malayang pag-guhit.
Maayo nga mosulat si Cesar.	Magaling na susulat si Cesar.	Magaling sumulat si Cesar.