Modeling Ilocano Enclitic Particles in Online Discourse as an Electronic Lexicon

Rillera, Hans
Saint Louis University
ABCR
Baguio City
2233195@slu.edu.ph

Roxas, Johan Rickardo
Saint Louis University
MRR-Queen of Peace
Baguio City
2233375@slu.edu.ph

Lacanilao, Marvin Patrick
Saint Louis University
Apugan-Loakan
Baguio City
2230153@slu.edu.ph

Jasmin, Ramon Emmiel P
Saint Louis University
Bakakeng Norte
Baguio City
2230748@slu.edu.ph

Tank, Rithik
Saint Louis University
Richwood Square Subd
Baguio City
2233293@slu.edu.ph

Siccuan, Sebastian
Saint Louis University
Trancoville
Baguio City
2233205@slu.edu.ph

De Torres, John Rey
Saint Louis University
Bakakeng Central
Baguio City
2234944@slu.edu.ph

ABSTRACT

Ilocano, a major Philippine language, employs a rich system of enclitic particles that shape discourse, convey speaker attitudes, and manage conversational flow. Despite their central role, these particles remain underrepresented in computational and descriptive resources. This study develops an electronic lexicon of Ilocano enclitic particles by integrating computational methods and expert linguistic judgment. Data were collected through web scraping and manual searches across platforms such as Reddit, YouTube, blogs, and online articles. Each instance of enclitic usage was systematically annotated and validated by native speakers, capturing morphological, semantic, and pragmatic information. The resulting lexicon organizes particles by form, category, contextual usage, and source, providing a structured resource that preserves both grammatical and discourse-level nuances. This work contributes to the documentation and computational modeling of Ilocano, offering a foundation for future natural language processing applications and discourse analysis.

CCS Concepts

Computing Methodologies → Artificial Intelligence →

Natural Language Processing → Language resources

Keywords

Enclitic particles, discourse particles, Ilocano, electronic lexicon

1. INTRODUCTION

1.1 Background of the Study

Enclitic particles are central to the pragmatics of natural language. In conversational English, expressions such as *you know*, *well*, and *like* guide interpretation rather than add propositional meaning [1]. These particles are often referred to as discourse particles, discourse markers, interjections, or fillers. At their core, however, they are pragmatic expressions with a procedural function: their contribution depends on the preceding word and conversational context rather than fixed lexical content [2]. They signal discourse relations, convey speaker attitude, and help manage the flow of interaction [3]. However, many linguists have dismissed them as "*coloring*" particles that should be desemanticized, claiming they hold no true meaning [4].

Ilocano, one of the major languages of the Philippines, features a particularly rich system of enclitic particles [5]. Common forms such as *-en*, *man*, *aya*, and *met* convey force, modality, orientation, or politeness [6]. Their prevalence and multifunctionality make them indispensable to Ilocano discourse.

1.2 Related Works

Research in computational linguistics has long grappled with the challenges of modeling discourse particles. Stede and Schmitz [7], for instance, examined German discourse particles in the context of discourse parsing, arguing that their contribution cannot be captured by syntax or semantics alone. They proposed treating them as pragmatic operators that signal discourse relations and guide interpretation, using rule-based methods that laid early groundwork for computational pragmatics.

In the Philippine context, Dita et al [8] pioneered the development of online corpora for Philippine languages, demonstrating how digital corpora can serve as foundational resources for linguistic description and computational modeling. Their work emphasizes the importance of building accessible electronic resources as a basis for further Natural Language Processing applications in under-resourced languages.

Moreover, Rubino's Ilocano Dictionary and Grammar [5] remains the most comprehensive descriptive resource, offering lexical documentation alongside grammatical notes on enclitics. However, Rubino did not develop a systematic lexical analysis of enclitic particles. Similarly, Schachter and Otanes [9], in their reference grammar of Tagalog, documented enclitic usage but treated particles primarily as grammatical markers rather than pragmatic resources. These works affirm the grammatical presence of enclitics but leave their discourse-level functions underexplored.

Building on these foundations, Balgos [6] provided a more focused analysis of Ilocano discourse particles. She classified them into four categories and demonstrated how they express opinions, attitudes, and emotions, thereby shaping discourse organization and meaning negotiation in interaction. This study highlights the central role of enclitic particles in Ilocano pragmatics and emphasizes the need for their systematic treatment as language resources.

1.3 Gap

Despite these contributions, discourse particles remain understudied in both descriptive and computational linguistics, often marginalized or treated as desemanticized elements. For Ilocano, Balgos' classification demonstrates their pragmatic richness, yet it has not been translated into resource development. Furthermore, while prior studies document enclitic particles in traditional written and spoken texts, their usage in online discourse — such as social media, blogs, and comment sections — remains largely unexplored. On these platforms, Ilocano speakers frequently employ particles in innovative ways, conveying pragmatic strategies, humor, emphasis, and interpersonal nuance specific to digital communication. This study addresses this gap by systematically collecting and analyzing data from online sources, capturing the dynamic, context-sensitive use of enclitic particles in contemporary Ilocano. To date, no electronic lexicon models these particles in terms of both lexical form and discourse function; developing such a resource is crucial, as these particles are indispensable for interpretation and interaction in Ilocano.

1.4 Objectives of the Study

This study proposes the development of an electronic lexicon for Ilocano enclitic particles. Specifically, it aims to:

1. Compile a comprehensive inventory of Ilocano enclitic particles from descriptive sources and digital text corpora.
2. Annotate and validate instances of enclitic particles in corpora to capture their pragmatic functions and syntactic behavior.
3. Construct an electronic lexicon that systematically organizes enclitic particles by form, category, usage context, and source information.

2. METHODOLOGY

2.1 Data Collection

The researchers performed several activities to achieve the objectives of the study which includes data collection, development of the lexicon, and data validation.

2.1.1 Webscraper

A custom webscraper was created to aid the researchers in collating resources needed for the creation of the lexicon. The webscraper was made in Python that scrapes for published content that are in the Ilocano or Filipino language that includes the usage of enclitic particles. It scraped for content on the platforms of Reddit, Wikipedia, YouTube comments, blogs, and other websites that do not have a login-first restriction. Each scraped content was automatically saved as a .txt file in the researchers' collaborative Google Drive with a unique identifier.

2.1.2 Manual Search

In addition to scraping websites, the study's proponents also opted to manually scour different platforms to collect data. In order to define some system in their approach, subgroups were deployed covering a specific platform. Because of the limitations presented with the webscraper, the data collected under this approach also proved pivotal with data ranging from varying sources including that of Youtube comments, Reddit comments, Facebook comments, blogs, and articles. Under each source, a minimum quota was set for the data to be collected and a key advantage experienced under this method was that a deeper level of analysis was conducted by the Ilocano language experts while they actively searched for viable data for the context of this study.

2.2 Data Annotation & Validation

After the collection of data through both web scraping and manual search, the gathered data underwent annotation and validation to ensure that the enclitic particles extracted were accurate and consistent. The researchers first collated the scraped and manually sourced content from websites, social media posts, blogs, online articles, and video transcripts. Instances containing enclitic particles were then isolated at the sentence or phrase level for further analysis.

The annotation process involved a systematic review of each entry. Sentences and phrases containing enclitic particles were segmented, and the particle in question was isolated for classification. Each particle was then categorized based on the functional groupings outlined in prior descriptive works, particularly Rubino [5] and Balgos [6]. In addition to classification, the annotation process recorded attributes such as the host word or phrase to which the enclitic attaches, its category, the corresponding English translation, the context, and source link.

Annotation and validation were conducted directly by the researchers who are themselves native or fluent speakers of Ilocano and process advanced familiarity with its grammar and pragmatics. Drawing on both linguistic expertise and cultural knowledge, the researchers cross-checked each annotation to ensure accuracy in form, categorization, and interpretation. They verified whether the identified word was indeed an enclitic particle, confirmed its assigned category, and refined the English translations to preserve pragmatic nuances. In cases where multiple interpretations were possible, consensus was reached through collaborative discussion. This integrated process of expert annotation and validation minimized subjectivity and strengthened the reliability of the gathered data.

The outcome of this stage was a rigorously annotated and validated corpus of Ilocano enclitic particles. By combining computational extraction with the researchers' expert linguistic judgment, the data not only captured the formal and pragmatic properties of enclitic particles but also ensured that their usage across different communicative contexts was represented. This annotated resource served as the foundation for the electronic lexicon.

2.3 Lexicon Creation

The creation of the electronic lexicon required the design of a structured schema to represent each instance of an Ilocano enclitic particle. The schema was built using insights from linguistic theory, grammatical frameworks, ontological principles, real corpus data, and existing lexical databases to ensure it accurately captures both the form and functions of enclitics.

The design drew heavily from the complex grammar of languages like Sankrit, which uses detailed grammatical analysis to handle affixes and word forms. It also applied principles that categorize words based on their meaning and relationships which help in clearly defining word properties. Furthermore, evidence from corpus linguistics was also incorporated to ensure that the entries reflect real-life usage, context-based meanings, and lexical semantic patterns. Finally , the schema integrates different ideas from established lexical databases such as WordNet and studies on classifiers in languages such as Nepali, Japanese and Chinese which helped in understanding how particles function and interact with words in phrases or sentences.

The lexicon is structured such that each row corresponds to a single annotated entry. To support both usability and interoperability, the schema is implemented in two formats: JSON and CSV , which contains the same set of fields, differing only in how the data is represented. These fields capture the linguistic, functional, and contextual properties of each particle, providing a structured resource for both descriptive analysis and computational processing. The schema contains the following fields:

id. A unique identifier for each entry.

word. The lexical host (word or phrase) to which the enclitic particle attaches.

enclitic_particle. The enclitic particle itself.

category. The pragmatic or functional category of the particle. Categories include:

Class 1: Completion, Addition

Class 2: Impatience / Command

Class 3a: Interrogative, Reporting

Class 3b: Pessimism / Emphasis

Class 4: Speculation

Class 5: Affirmation

Other: Speculation, Hope / Necessity, Discovery

snippet_usage. A sentence or phrase excerpt showing authentic usage.

english_translation. English rendering of the snippet, preserving both literal meaning and pragmatic nuance.

context. The discourse environment or conversational situation of occurence

pragmatic_function. Communication effect that gives deeper meaning to the context

clictic_position. Placement of a clitic relative to words in the sentence or phrase

notes. Supplementary remarks

source_url. The online source for transparency and verification

This schema provides the structural foundation for the electronic lexicon. It ensures that each entry is recorded in a clear, consistent, and organized way. With this structure, the lexicon becomes easier to analyze, expand, and utilize for both linguistic research and computational applications.

3. FINDINGS

3.1 Data Collection

The data collection process combined automated and manual methods to compile a diverse corpus of Ilocano enclitic particles. A custom Python-based web scraper extracted textual data from publicly accessible sources, including Reddit, Wikipedia, YouTube comments, blogs, and media websites. Each piece of content was saved with a unique identifier for traceability.

In addition to automated scraping, the research team conducted targeted manual searches to supplement the dataset. Subgroups of researchers focused on specific platforms, collecting data from sources that were inaccessible or limited for the scraper. This approach allowed for nuanced selection, ensuring the inclusion of contextually rich examples of enclitic usage. A minimum data quota per source was maintained to guarantee dataset breadth.

Through this combined strategy, the researchers collated 135 digital resources capturing authentic instances of Ilocano enclitic particles. The dataset provides a representative foundation for subsequent annotation, validation, and lexicon development, reflecting both formal grammatical structures and real-world communicative contexts.

3.2 Data Annotation & Validation

Each lexical entry was carefully annotated by the researchers. These annotations were carried out by the Ilocano experts within the team. For every entry, the morphological features including the enclitic particle and clitic position, semantic category, and English translation were recorded. Also, other contextual notes were made like the pragmatic function.

To ensure the accuracy of the data annotation, the process followed a structured set of guidelines:

Morphological annotation. Each annotation was checked against existing literature on grammar and translation. Also, it was verified by native speakers of the Ilocano language.

Semantic categorization. The category of the words were applied consistently across different entries using manual annotation which was cross-validated using the transformer-based large language model.

Pragmatic Functions. The functions were defined by examining natural usage in texts and a group deliberation was conducted to verify the annotations of the group.

The overall validation process was carried out by means of the triangulation process. This process covered a three-step approach that covered combining the judgement of native speakers of the Ilocano language, referencing existing literature, and through internal consistency checks covering the proper formatting of the data input.

3.3 Lexicon Creation

In total, 135 lexical entries were considered in creating the lexicon. Each entry was structured into a standardized format, recording not only the base word or phrase but also its associated morphological, semantic, and pragmatic information. The entries followed a consistent template including: the word/phrase, enclitic particle, semantic category, snippet usage, English translation, contextual notes, pragmatic function, clitic position, and source reference. This structure ensured that each entry could stand alone as an informative unit while still aligning with the overall system of the lexicon.

An example is presented in Table 1, which illustrates the entry for adda with the enclitic particle met.

Table 1. Example of lexicon entry (*adda met*)

JSON

```json
{ "id": 1,
"Word/Phrase": "adda",
"enclitic_particle": "met",
"category": "Class 3b: Pessimism / Emphasis",
"snippet_usage": "\"Adda met dagiti taraken da Pader.
Pabo, kuneho, puraw a bao a babassit, aso, pusa…"",
"english_translation": "\"Father also has livestock:
turkey, rabbit, a small white mouse, dog, cat…"",
"context": "Met adds contrast, \"but/also\" nuance.",
"pragmatic_function": "Contrastive/additive particle
('also, too, but') that softens assertion and manages
discourse flow.",
"clitic_position": "Enclitic; attaches after the first
stressed word of the clause (often the predicate or
topicalized element).",
"notes": "",
"source_url":
"https://arielsotelotabag.blogspot.com/2008/09/littugaw.
html" }
```

The phrase, *"Adda met dagiti taraken da Pader. Pabo, kuneho, puraw a bao a babassit, aso, pusa…"*, illustrates the usage of the enclitic *met* in shaping clause-level meaning. Here, the word *adda* does not simply indicate existence (as its English equivalent "there is/are" would suggest), but is modified by *met* to convey an additive nuance, roughly equivalent to "also" or "but" in English. This highlights how the lexicon captures not only the literal meaning of a word but also the pragmatic and discourse functions that emerge in real usage.

By systematically compiling such entries, the lexicon provides a resource that reflects both the grammatical structure and the communicative subtleties of Ilocano. This dual emphasis makes it useful not only for linguistic analysis but also for applications in language preservation, pedagogy, and computational modeling.

4. CONCLUSION

This study set out to address the gap in Ilocano linguistic resources by developing an electronic lexicon of enclitics particles. Through a combination of computational methods, such as web scraping, and manual collection assisted by native speakers, the researchers compiled a diverse corpus of enclitic usage drawn from digital and conversational contexts. The subsequent data then underwent rigorous annotation and validation to ensure that the pragmatic and functional nuances of these particles were preserved, allowing the lexicon to capture not only their lexical forms but also their discourse-level roles.

The creation of the lexicon demonstrates the potential of integrating descriptive linguistics with computational approaches for under-resourced languages. By organizing Ilocano enclitics into a structured schema, this study provides a foundation that can support both linguistic analysis and future Natural Language Processing applications. In doing so, the study underscores the pragmatic richness of Ilocano and establishes discourse particles as indispensable elements in shaping interpretation, guiding interaction and constructing meaning with communication.

For future studies, it is recommended to extend the lexicon by incorporating pragmatic rules and discourse functions of enclitic particles, particularly regarding precedence and placement in grammar, as suggested by Balgos. Such an extension would allow the lexicon to function not only as a lexical inventory but also as a pragmatic resource for discourse-level analysis. Moreover, computational applications such as classifiers for automatic recognition of enclitic particles or dialogue systems that integrate enclitic usage could enhance naturalistic interaction, especially in social media and other online discourse contexts.

5. REFERENCES

[1] Schoroup, L. C. 1983. Common Discourse Particles in English Conversation. In *Working Papers in Linguistics*, 28(1). https://doi.org/10.4324/9781315401584

[2] Fraser, B. 1999. What are discourse markers? In J*ournal of Pragmatics 31*(7). 931-952. https://doi.org/10.1016/S0378-2166(98)00101-5

[3] Schiffrin, D. 2005. Discourse Markers: Language, Meaning, and Context. In *The Handbook of Discourse Analysis*. https://doi.org/10.1002/9780470753460.ch4

[4] Skettekorn, W. 1977. Pragmatics and Discursive Rhetoric. Journal of Pragmatics.

[5] Rubino, C. R. G. 1997. A Reference Grammar of Ilocano. *University of California*.

[6] Balgos, A. R. G. 2019. '*The particle is the meaning, aya?': Discourse Particles in Ilocano*. UP Diliman Review. https://www.humanitiesdiliman.upd.edu.ph/index.php/dilimanreview/article/view/10252/9077

[7] Stede, M. & Schmitz, B. 2000. Discourse Particles and Discourse Functions. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000). Association for Computational Linguistics, Saarbrücken, Germany*, 800–806. https://doi.org/10.1023/A:1011112031877

[8] Dita, S. N., Roxas, R. E. O. & Inventado, P. 2009. Building Online Corpora of Philippine Languages. In *23rd Pacific Asia Conference on Language, Information and Computation.* 646-653. https://animorepository.dlsu.edu.ph/faculty_research/2952

[9] Schachter, P. & Otanes, F. T. 1972. Tagalog Reference Grammar. *Berkeley: University of California Pres.* https://doi.org/10.1177/00336882730040011