Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3270~3276
e-ISSN: 2548-6861

Construction of a Dialect-Sensitive Javanese Semantic Lexicon to Support Machine Translation Systems

Musthofa Galih Pradana [1]*, Ridwan Raafi'udin[2], Nurul Afifah Arifuddin[3]**, Mohammad Asaduzzaman Rasel[4]*****
* Data Science Department, Computer Science Faculty, Universitas Pembangunan Nasional Veteran Jakarta, Indonesia
** Informatics Department, Computer Science Faculty, Universitas Pembangunan Nasional Veteran Jakarta, Indonesia
*** School of Information Technology, Monash University, Malaysia
musthofagalihpradana@upnvj.ac.id [1], raafiudin@upnvj.ac.id [2], nurulafifaharifuddin@upnvj.ac.id [3],
mohammadasaduzzaman.rasel@monash.edu [4]

Article Info

Article history:

Received 2026-06-18
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Javanese Semantic Lexicon,
Speech Level Variation,
Lexical Resources,
Low-Resources Languages,
Polysemy Exploration.

ABSTRACT

The development of linguistic resources for natural language processing (NLP) in Javanese remains limited, especially regarding the representation of semantic relationships between different speech levels. This study aims to construct a Javanese semantic lexicon that integrates Indonesian lexical equivalents with three Javanese speech levels: ngoko, krama alus, and krama inggil. A research design based on lexical resource construction was employed, using a Javanese digital dictionary as the primary data source. The methodology included data extraction, preprocessing, semantic lexicon construction, analysis of speech level variation, and a preliminary exploration of polysemous lexical entries using automatic identification, followed by validation by native speakers. The resulting semantic lexicon successfully represents lexical relationships between levels in a structured manner. Analysis of speech-level variation revealed that partially distinct lexical patterns were the most dominant, with 733 entries, followed by fully distinct patterns (193 entries) and identical patterns (21 entries). These findings indicate that speech-level differences in Javanese are selectively realized and should be explicitly considered in the development of linguistic resources. Furthermore, preliminary exploration of polysemous candidates demonstrated that dictionary-based automatic identification can overestimate polysemy without linguistic validation. Only a limited number of lexical entries exhibited features consistent with genuine polysemous relationships. This study provides an initial basis for the development of Javanese semantic resources that are sensitive to speech-level variation and semantic complexity. The constructed semantic lexicon has the potential to support future research in NLP applications in Javanese, including politeness identification, lexical normalization, word sense disambiguation, and machine translation.

I. INTRODUCTION

Indonesia is a country with a diverse culture and language. One of the regional languages in Indonesia is Javanese, which is the language with the largest number of speakers in Indonesia and plays a significant role in the socio-cultural life of the community. In addition to its large number of speakers, Javanese also has complex linguistic characteristics, especially in its implementation in daily life due to the variety of language levels to describe the conditions of social relations between speakers, such as the levels of ngoko, krama alus, and krama inggil. This is an added value of Javanese, which is its distinctive and unique characteristics. This is just one of the complexities of Javanese, which certainly poses challenges in the development of natural language processing technology for Javanese. Research in the field of natural language processing, especially machine translation for regional languages categories, is categorized as low-resource languages, experiencing significant development and interest. Approaches such as Neural Machine Translation show good

performance in languages that have the availability of large parallel corpuses [1]. However, in low-resource languages, this remains a major obstacle. Most previous research has focused on building parallel corpora and developing machine translation, without considering the varying levels of speech that can represent the social relationships between speakers. This can lead to translation difficulties and ambiguity.

The aforementioned issues are reinforced by relevant previous research, such as the Javanese language's speech levels, which include ngoko (speech level), krama alus (good manners), and krama inggil (high manners). Relevant research shows that modern language models struggle to accurately understand the meaning and honorific forms of Javanese due to its high complexity [2]. This indicates that lexical representation remains a challenge in machine translation. While machine translation approaches, particularly those using Neural Machine Translation, perform well, they still face challenges and constraints in low-resource languages, especially those requiring parallel corpus availability [3]. Discussing further in the Javanese-Indonesian context, many studies focus on improving the performance of translation models through optimization without paying attention to the semantic aspects and variations in speech levels that characterize Javanese [4].

Another problem that arises in Javanese machine translation besides the speech level is the existence of the phenomenon of polysemy, namely a condition where a lexical form has more than one meaning which can trigger ambiguity in the meaning of the translation [5]. This polysemy problem was conveyed by several previous researchers, such as the integration of semantic information and word sense disambiguation mechanisms are still needed to improve modern translation [6]. The availability of a semantic lexicon model that can understand and represent the relationship between concepts at the speech level in Javanese has not been explored further [7]. Semantic lexicon has great potential in the large task of natural language processing, especially in machine translation, because it has a more sensitive character to complex linguistic details such as Javanese.

Previous research exploring the complexity of the Javanese honorific system has begun to increase. The use of model optimization to identify Javanese speech levels has been shown to improve translation performance, such as improved performance with the enhanced M2M100, achieving optimal results [8]. Other optimizations are also carried out using pivot-based techniques, although they still have limitations in the form of errors and biases arising from the pivot language [9]. The solution to this problem by using neural machine translation indeed shows the selection of the right method with significant results obtained such as the BLEU score increasing from 15% to 81% after refining the model and dataset with the Javanese language configuration by paying attention to the speech level aspect [10] compared to machine translation which still uses natural machine translation such as Levenshtein distance [11].

Moving on to the next issue regarding ambiguity that arises due to the phenomenon of polysemy, relevant research shows that handling ambiguity in Indonesian language reviews that with part of speech (POS) marking, word feature extraction and semantic measurement can be done and studies on this are challenging with the right methodology [12]. Handling polysemy can be done using a hybrid approach, namely a combination of lexical resources with a transformer-based model. This finding illustrates the importance of developing structured semantic resources [13]. Similar to the problem of dialect variation using the pivot mechanism, the previous application for polysemy is also the same. By integrating NER and pivoting using NMT, experimental results show that translation accuracy fluctuates between 0.32 and 0.54 for Javanese to Madurese and 0.21 to 0.45 for Madurese to Javanese [14]. Another thing is that to improve NMT in overcoming ambiguity problems, advanced language processing methodologies are needed as well as additional linguistic resources [15]. In addition, other findings show that the resolution of polysemy is more optimal if it is based on rule-based translation compared to neural machine translation [16]. Another alternative is to use a hybrid rule-based and neural machine translation.

The results of the study and previous research references, it can be identified that the emerging research gaps are: (1) Most of the low resource language research, especially Javanese, is still oriented towards model development, not yet touching the realm of semantic resources. (2) There are still limited semantic resources in Javanese that explicitly accommodate and understand the context of speech levels such as ngoko, krama alus, and krama inggil. (3) Semantic phenomena that occur such as polysemy and synonyms have not been explored further which should be an integral part of the development of Javanese linguistic resources. In order to resolve the research gap, this study proposes an initial exploration in the identification of semantic lexicons from Javanese to Indonesian that has taken into account the sensitivity of variations in speech levels orchestrated from digital dictionary data sources. In addition, this study will also model the initial mitigation of polysemy and synonym candidates that appear in Javanese.

The main contributions highlighted in this research are in accordance with the 3 main problems that have been conveyed in the research gap point, namely: (1) The research produces a semantic lexicon model of Javanese-Indonesian with conditions in low-resource languages. (2) Linguistic analysis and models with consideration of variations in dialectical speech levels. (3) Initial exploration of the existing dataset for semantic phenomena, namely polysemy. So it is hoped that this model can be the basis for the development of machine translation technology that accommodates various types of problems that are the focus of research in the field of machine translation. Thus, there is still a lack of dialect-sensitive semantic lexicons in Javanese, with limited resources. This research will propose the development of a semantic lexicon with a level of dialect sensitivity that reflects the existing level of Javanese speech.

II. METHOD

The method section of this research will describe the flow of the research conducted starting from data collection, information regarding data description, data processing, as well as the stages and analysis needs that will be used to solve the problems that have been stated in the points in the research gap analysis in the previous introduction.

A. Research Framework

This research applies a language resource construction approach to construct a semantic lexicon from Javanese to Indonesian. The development is based on consideration of dialectical speech level variations as a key characteristic. The research framework also includes analysis and construction of an initial semantic-based model by identifying polysemy phenomena frequently encountered in Javanese. This complete framework will produce a model capable of learning by considering dialect variations based on low-resource language and initial identification of polysemy models with detailed flow show on Figure 1.

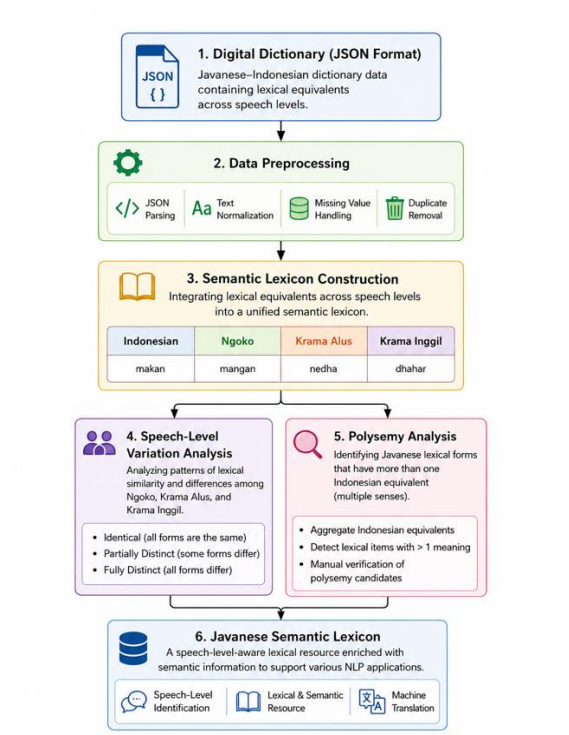

Figure 1. Research Stage.

B. Dataset Description

The dataset used in this research is taken from a Javanese to Indonesian digital dictionary in JSON format. The dataset contains lexical representations of Indonesian and Javanese based on variations in dialect levels that are characteristic of Javanese. A general description of the dataset has the following attributes: (1) Indonesian. (2) Ngoko. (3) Krama Alus. (4) Krama Inggil. These are levels commonly used to represent variations in dialect levels that refer to aspects of social communication between speakers. An overview of the dataset description is shown in Table *I*.

Table I. Dataset Description

Indonesian	Ngoko	Krama Alus	Krama Inggil
Makan	Mangan	Nedha	Dhahar
Tidur	Turu	Tilem	Sare
Pulang	Mulih	Kondur	Kondur
Melihat	Ndeleng	Ningali	Mirsani
Berkata	Ngomong	Matur	Ngendika

The dataset used in the semantic construction process must first go through a pre-processing stage. Because the data format is JSON, the stages that need to be carried out include: JSON parsing, text normalization, handling empty entries, and removing duplicate data. This process will be explained in more detail in the next sub-chapter.

C. Data Pre-Processing

As mentioned in general in the previous point, data pre-processing is necessary, the main goal of which is to improve data quality and consistency before the data is further processed in the semantic lexicon construction. The pre-processing steps are carried out according to the procedures in TABLE II.

TABLE II.
DATA PRE-PROCESSING PROCEDURES

Step	Purpose	Procedures
JSON Parsing	Convert dictionary entries into tabular format	JSON Data Frame
Text Normalization	Ensure consistent lexical representation	Lowercasing and whitespace trimming
Missing Value Handling	Remove incomplete entries	Remove records without Indonesian or Ngoko forms
Duplicate Removal	Eliminate redundant lexical pairs	Retain unique entries only

After the procedure is completed, the next step is a systematic data preprocessing workflow. The main steps begin with JSON parsing, which aims to convert the digital dictionary data structure to tabular format. This is followed by text normalization, which involves lower casing and whitespace trimming. Next, the data is ensured to be meaningful by handling empty values. Finally, duplicate data is removed to ensure data consistency. The complete flow is shown in Figure *2*.

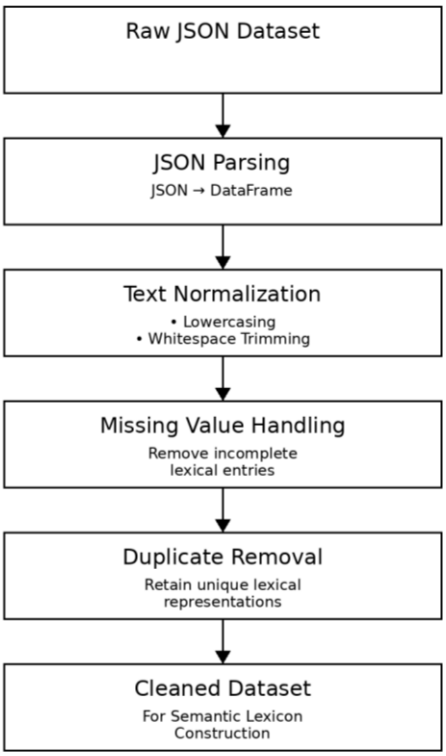

Figure 2. Pre-Processing Stage

D. Semantic Lexicon Construction

The next step after pre-processing is to construct a semantic lexicon. This study defines a semantic lexicon as a linguistic resource capable of describing relationships between concepts through lexical equivalents and their speech-level structures. Furthermore, this section constructs these four attributes into a unified lexical structure. The semantic construction stages involve the following mechanisms: (1) Concept extraction from the dataset. (2) Integration of speech-level lexical equivalents. (3) Formation of the semantic lexicon structure. The semantic lexicon construction process consists of the following five stages:

Step 1. Identification of Lexical Concepts

Existing lexical items are implemented as concept nodes because they provide a language-independent representation. All entries are normalized in pre-processing and grouped according to their Indonesian lexical meaning.

Step 2. Lexical Alignment at the Speech Level

Relevant and appropriate Javanese lexical forms are aligned at the speech level used, namely ngoko, krama alus, and krama inggil. This represents a single semantic structure while maintaining distinctions related to politeness levels and social hierarchy.

Step 3. Semantic Relationship Construction

Semantic relationships are established by connecting Indonesian concept nodes with lexical realizations at the speech level. This allows the semantic lexicon to capture lexical equivalence while maintaining the speech level distinctions found in conventional dictionary data corpuses.

Step 4. Classification of Speech-Level Patterns

Lexical variation can be analyzed within each semantic concept and falls into one of three lexical patterns: Identical, when all conditions are exactly the same at the speech level. Partially distinct, when only some speech levels exhibit lexical variation. Fully distinct, when each speech level uses a different lexical realization.

Step 5. Initial Polysemy Identification

In the final stage, the constructed semantic lexicon is identified for potentially polysemous lexical entries. The identification process is carried out automatically, generating polysemy candidates. These will be validated by native Javanese speakers to re-validate the consistency of the automatically generated data.

III. RESULT AND DISCUSSION

The semantic lexicon construction process described previously yields Javanese-Indonesian resources according to speech level variations. This data contains Javanese-Indonesian equivalents at the ngoko, krama alus, and krama inggil speech levels. This structure allows for representation in various lexical forms according to social context.

A. Speech-Level Variation Analysis

Analisis variasi tingkat tutur yang dibangun digunakan to map and see the lexical relationship patterns between speech levels such as ngoko, krama alus, krama inggil in a semantic lexicon. The grouping will be done into identical, partially distinct, and fully distinct. (1) The identical category shows that the concept has the same lexical representation at the speech level. The indication is that changes in politeness levels are not always followed by changes in lexical form. (2) The fully distinct category shows that each speech level has a different lexical form. Word choice plays an important role here, to determine the social context and level of politeness. (3) The partially distinct category shows a condition between the two patterns, when only a portion of the speech level experiences changes in lexical form. The distribution of speech levels obtained in the data of this study is shown in Figure *3*.

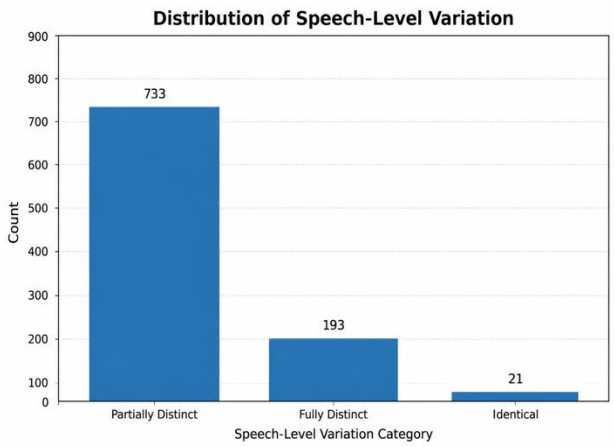

Figure 3. Distribution of Speech Level

Based on the data distribution, the highest percentage of partially distinct words, with 773 words or 77.4%, indicates that changes in the Javanese language's speech level generally only affect some lexical forms. The fully distinct category, with 193 words representing 20.4%, indicates that concepts undergo comprehensive lexical changes across speech levels. Finally, the 21 words or 2.2% use of the same lexical form across all speech levels, a relatively rare occurrence.

B. Preliminary Exploration of Polysemous Lexical Entries

Following the analysis of speech-level variation, this study attempted to conduct an initial exploration of the potential for polysemy in the constructed semantic lexicon. This analysis aimed to identify Javanese lexical forms that have the potential for more than one meaning representation in Indonesian. This polysemy phenomenon is important to understand more deeply, as it can influence the quality and process of machine translation lexical equivalents.

These candidates are ngoko lemmas mapped to more than one Indonesian equivalent in the dataset. Referring back to the established initial procedure, this experiment yielded 19 polysemy candidates. These polysemy candidates were obtained by searching for ngoko lemmas that have more than one Indonesian equivalent in the constructed semantic lexicon. The findings are shown in TABLE III.

TABLE III.
POLYSEMY CANDIDATES

Number	Ngoko	Number of Meaning
1	Akon	2
2	Arep	2
3	Awak	2
4	Ayo	2
5	Becik	2
6	Deleh	2
7	Dhemen	2
8	Eling	2
9	Genti	2
10	Gering	2
11	Iwak	2
12	Katekan	2
13	Kepenak	2
14	Nagjokake	2
15	Ngaku	2
16	Oleh	2
17	Pacoban	2
18	Padusan	2
19	Semana	2

These data provide examples of polysemy candidates found in the dataset in this study. However, further analysis and validation are needed, and these results should not be interpreted as valid. Before conducting a more in-depth analysis, here are detailed data on the word equivalents generated by the experimental procedure on this dataset, as shown in TABLE IV.

TABLE IV.
PAIRED MEANING POLYSEMY CANDIDATES

Ngoko	Indonesian Meaning 1	Indonesian Meaning 2
Akon	Dianggap	Disuruh
Arep	Akan	Ingin
Awak	Badan	Tubuh
Ayo	Mari	Silahkan
Becik	Baik	Bagus
Deleh	Taruh	Letak
Dhemen	Senang	Suka
Eling	Ingat	Sadar
Genti	Gilir	Ganti
Gering	Sakit	Kurus
Iwak	Ikan	Daging
Katekan	Kedatangan	Kesampaian
Kepenak	Perasaan enak	Sembuh
Nagjokake	Mengajukan	Mencalonkan
Ngaku	Mengaku	Mengakui
Oleh	Dapat	Memperoleh
Pacoban	Percobaan	Cobaan
Padusan	Tempat Mandi	Mandi Sebelum Puasa
Semana	Sekian	Demikian

The results show that the polysemy data generated still contains misrepresentations of the concept of polysemy. These results were validated by native Javanese speakers, who stated that most of the candidates were found as causal synonyms in Indonesian, not purely polysemous meanings in Javanese. The problem arises from lexicographic inconsistencies and variations in conceptual interpretation. An example of a fully valid polysemous word is the word "gering," which has two meanings: "thin" and "sick." Furthermore, the meaning of "iwak," both for "fish" and "meat," was also validated as a valid polysemy, with the remaining errors resulting from the dataset.

Native speakers were selected based on the following criteria: (1) Native speakers of Javanese who use the language in everyday communication. (2) Familiarity with the three standard speech levels (ngoko, krama alus, and krama inggil). (3) Sufficient linguistic competence to distinguish lexical variation, semantic equivalence, and contextual word usage. During the validation process, each lexical entry was

independently reviewed by validators. For each semantic concept, the validators checked whether the lexical correspondences between Indonesian, ngoko, krama alus, and krama inggil accurately represented the intended meaning. For automatically detected polysemy candidates, the validators determined whether the Indonesian equivalents reflected true polysemy or simply synonyms, lexical variations, or inconsistencies in dictionary construction. Any discrepancies identified during the validation process were resolved through discussion until consensus was reached. The validated entries were then incorporated into the final semantic lexicon used for analysis.

The findings of this study indicate that the identification of polysemy from Javanese to Indonesian remains very limited without adequate linguistic validation. Therefore, the authors position this as an initial exploration that will lay the foundation for future research and dataset development in the context of Javanese polysemy by enriching and validating dictionary data that will be used for machine translation models, understanding the context at the speech level.

C. Implications for Javanese NLP Resources

This study identified several verified potential polysemy instances, albeit in small numbers. These results suggest that there is still a lack of linguistic resources proportional to the needs and challenges of machine translation in the context of accommodating polysemy phenomena. The results also indicate that the development of a semantic lexicon capable of considering both the speech level and polysemy phenomena is crucial for further development. The analysis of speech level variation also provides important insights, as differences in speech level in Javanese do not always involve changes in the overall lexical form, but only occur for certain concepts that are sensitive to social context and politeness levels. Therefore, an important implication is that speech level information potentially misses important aspects of the Javanese linguistic system if lexical representation ignores speech level information.

Initial exploration of candidate polysemy data indicates that the semantic lexicon can be utilized to identify ambiguity phenomena that arise in Javanese, although further validation and verification are still needed. These findings demonstrate that the semantic lexicon has great potential as a supporting resource in research related to meaning representation in low-resource languages. The main contribution of this research lies not only in the compilation and identification of resources on various computational tasks such as speech level, politeness level and selection of lexical equivalents.

Although this study primarily focuses on semantic resource construction, the resulting dialect-sensitive semantic lexicon provides a practical foundation for several downstream natural language processing applications, particularly machine translation for resource-constrained Javanese. The semantic lexicon is designed to complement data-driven translation models by providing structured lexical knowledge typically unavailable in parallel corpora. First, the semantic lexicon can improve lexical selection during translation. Conventional neural machine translation systems typically learn lexical correspondences from statistical patterns in parallel corpora. However, because Javanese exhibits multiple lexical realizations depending on the speech level, translation models often produce semantically correct but pragmatically inappropriate translations. The proposed semantic lexicon explicitly represents lexical alternatives for ngoko, krama alus, and krama inggil, enabling future translation systems to select lexical forms that better preserve politeness and social context. Second, the semantic lexicon can support word-meaning disambiguation (WSD). Initial identification and native speaker validation of polysemic lexical entries indicate that some Javanese lexical forms correspond to multiple semantic concepts. Integrating this lexical knowledge into a translation system could aid in selecting context-appropriate word meanings and reduce semantic ambiguity during translation. Third, the semantic lexicon could serve as an external lexical knowledge base for retrieval-enriched or lexically constrained neural machine translation. Instead of relying solely on learned parameters, the translation model could consult the semantic lexicon to retrieve candidate lexical forms before producing the final translation. Such integration is expected to improve lexical consistency while preserving the speech-level distinctions important in Javanese communication.

D. Limitations of the Study

The limitations of the research revealed need to be identified and become insights for the continuation of further research. (1) The semantic lexicon constructed is derived from only one digital dictionary, so the coverage does not fully represent all variations in the use of Javanese. (2) The analysis of speech level variations applied in the research is only focused on the representation of ngoko, krama alus, and krama inggil. This variation can be influenced by geographical factors, such as the Banyumasan dialect, East Java, which has a different character and has its own unique characteristics. (3) Initial exploration of the polysemy candidate data was carried out using an approach based on lexical correspondence in the semantic lexicon. In this research, direct validation was also carried out by native speakers, but this process still does not involve annotation by linguists or corpus evaluation. (4) This research is also still limited in the evaluation stage carried out. The evaluation carried out has not directly evaluated the application of the resulting semantic lexicon, therefore the effectiveness of the semantic lexicon model still needs to be further proven in further research.

The primary objective of this research is to lay the foundation for more comprehensive semantic lexicon model research. Therefore, this research does not focus on measuring predictive accuracy, a common practice in previous research. Instead, it focuses on generating relevant validation for the development of machine translation tools that have a rich

understanding and representation of the Javanese language and all its characteristics. Comprehensive quantitative evaluation, including assessment of intrinsic lexical quality and downstream evaluation in machine translation systems or word meaning disambiguation, remains an important direction for future research.

IV. CONCLUSION

The research conducted has successfully constructed a semantic lexicon for Javanese and Indonesian by integrating across speech levels, namely ngoko, krama alus, and krama inggil. Compared to conventional bilingual dictionaries, the resulting semantic lexicon not only matches lexical words but also maintains the relationship between speech levels as part of the semantic representation of Javanese, which is a unique value and main characteristic. The research findings also show that variations in speech levels are a prominent characteristic of Javanese lexical resources, as indicated by the partial variations found, where changes in politeness levels do not necessarily always occur in the replacement of all lexical forms. Another finding that answers the research question regarding the initial exploration of candidate polysemy phenomena indicates that dictionary-based identification has the potential to produce overtranslation. Therefore, manual validation by speakers is necessary to optimize the results obtained from the generated model. The validation results indicate that the polysemy discovered is mostly due to synonymy in Indonesian, not a polysemy phenomenon.

These findings suggest a direction for further research: a Javanese semantic lexicon requires more structured data sensitive to local language characteristics. Developing adaptive and sensitive resources is a challenge that must be addressed to support further research in the field of machine translation.

ACKNOWLEDGEMENT

The authors declare no conflict of interest. This research was funded by the Institute for Research and Community Service (LPPM), Universitas Pembangunan Nasional Veteran Jakarta, through the International Collaborative Research (RIKIN) scheme in 2026.

REFERENCES

[1] A. Z. Munibi, "Assessing Neural Machine Translation in Speech: Problems and Solutions in AI-Powered Translations," *Journal of Computing Innovations and Emerging Technologies*, vol. 2, no. 1, pp. 25–32, Jun. 2026, doi: https://doi.org/10.64472/jciet.v2i1.27

[2] M. R. Farhansyah, I. Darmawan, A. Kusumawardhana, G. I. Winata, A. F. Aji, and D. T. Wijaya, "Do Language Models Understand Honorific Systems in Javanese?," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 26732–26754. doi: https://doi.org/10.18653/v1/2025.acl-long.1296

[3] W. Tan and K. Zhu, "NusaMT-7B: Machine Translation for Low-Resource Indonesian Languages with Large Language Models," Oct. 2024. doi : https://doi.org/10.48550

[4] F. I. Putri, A. P. Wibawa, and L. H. Collante, "Refining the Performance of Indonesian-Javanese Bilingual Neural Machine Translation Using Adam Optimizer," *ILKOM Jurnal Ilmiah*, vol. 16, no. 3, pp. 271–282, Dec. 2024, doi: https://doi.org/10.33096/ilkom.v16i3.2467.271-282

[5] E. Rippeth, M. Carpuat, K. Duh, and M. Post, "Improving Word Sense Disambiguation in Neural Machine Translation with Salient Document Context," Nov. 2023. doi : https://doi.org/10.18653/v1/2024.emnlp-demo.34

[6] H. D. Masethe, M. A. Masethe, S. O. Ojo, P. A. Owolawi, and F. Giunchiglia, "Hybrid Transformer-Based Large Language Models for Word Sense Disambiguation in the Low-Resource Sesotho sa Leboa Language," *Applied Sciences*, vol. 15, no. 7, p. 3608, Mar. 2025, doi: https://doi.org/10.3390/app15073608

[7] R. Goworek *et al.*, "SenWiCh: Sense-Annotation of Low-Resource Languages for WiC using Hybrid Methods," in *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 61–74. doi: https://doi.org/10.18653/v1/2025.sigtyp-1.7

[8] M. B. Prayoga, B. R. Ermawan, A. R. Fadhillah, M. N. Farizi, and E. Utami, "Evaluating Machine Translation Models and LLMs for Indonesian–Javanese Translation Across Speech Levels," *Journal of Applied Informatics and Computing*, vol. 10, no. 2, pp. 1549–1560, Apr. 2026, doi: https://doi.org/10.30871/jaic.v10i2.12326

[9] D. A. Sulistyo, A. P. Wibawa, D. D. Prasetya, and F. A. Ahda, "An enhanced pivot-based neural machine translation for low-resource languages," *International Journal of Advances in Intelligent Informatics*, vol. 11, no. 2, p. 258, May 2025, doi: https://doi.org/10.26555/ijain.v11i2.2115

[10] C. I. Ratnasari and D. D. Ramadhan, "Neural Machine Translation from Indonesian to Javanese Ngoko using RNN-LSTM Approach," in *2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, IEEE, Feb. 2026, pp. 1–6. doi: https://doi.org/10.1109/ACDSA67686.2026.11468163

[11] M. G. Pradana, H. B. Seta, N. Irzavika, P. H. Saputro, and R. Rusiyono, "Levenshtein Distance Algorithm in Javanese Character Translation Machine Based on Optical Character Recognition," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 4, p. 1411, Jul. 2025, doi: https://doi.org/10.62527/joiv.9.4.3151

[12] A. Nurhadiyatna, D. Rahman, and T. Hidayat, "Improving Javanese OCR via Transliteration and Levenshtein-based Post-Processing," *Journal of Intelligent Informatics and Visualization (JOIV)*, vol. 4, no. 2, pp. 89–98, 2023, doi: https://doi.org/10.35735/joiv.v4i2.1579

[13] H. D. Masethe, M. A. Masethe, S. O. Ojo, P. A. Owolawi, and F. Giunchiglia, "Hybrid Transformer-Based Large Language Models for Word Sense Disambiguation in the Low-Resource Sesotho sa Leboa Language," *Applied Sciences*, vol. 15, no. 7, p. 3608, Mar. 2025, doi: https://doi.org/10.3390/app15073608

[14] D. A. Sulistyo, D. D. Prasetya, F. A. Ahda, and A. P. Wibawa, "Pivoted Low Resource Multilingual Translation with NER Optimization," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 5, pp. 1–16, May 2025, doi: https://doi.org/10.1145/3727876

[15] F. Almu'iini Ahda, A. Prasetya Wibawa, D. Dwi Prasetya, D. Arbian Sulistyo, and A. Nafalski, "Advanced Machine Translation-Based Stammer for Preserving the Minangkabau Traditional Medicine Domain," *JOIV : International Journal on Informatics Visualization*, vol. 10, no. 3, p. 959, May 2026, doi: https://doi.org/10.62527/joiv.10.3.4042

[16] "Word Sense Disambiguation and Location Named Entity Recognition for Madurese-Indonesian Rule-based Machine Translation," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 2, pp. 572–588, Mar. 2025, doi: https://doi.org/10.22266/ijies2025.0331.42