JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

Building a Digital Lexical Resource for Banyumasan Javanese: A Low-Resource Language Approach

Nisrina Hanifa Setiono[1*], Angga Kurniawan[2], Viga Laksa Hardjanto[3]
[1,2,] Data Science, Telkom University Purwokerto, Indonesia
[3] Informatics Engineering, BINUS Online Learning, Bina Nusantara University, Indonesia
[1*]nisrinahanifas@telkomuniversity.ac.id, [2]anggakurniawan@telkomuniversity.ac.id,
[3] viga.hardjanto@binus.ac.id

Abstract
Banyumasan Javanese, widely recognized through the Ngapak dialect, remains culturally significant but is still underrepresented in reusable computational resources. This study develops a digital Banyumasan-Indonesian lexical corpus and frames it as a reusable research artifact rather than a static appendix. The corpus was constructed from a Banyumasan-Indonesian dictionary, normalized into a structured bilingual dataset, and packaged as an installable Python resource so that it can be used directly in computational experiments. The implemented resource supports dataset loading, Banyumasan lookup, Indonesian lookup, simple translation, structured translation analysis, batch translation, and corpus statistics. The resulting corpus contains 2,000 lexical pairs, 1,996 unique Banyumasan forms, 1,444 unique Indonesian equivalents, and 4 duplicated Banyumasan headwords that preserve lexical ambiguity from the source material. To demonstrate practical utility, the study includes a 100-sentence implementation example in which Banyumasan text is translated with the published banyumasan-corpus package and evaluated against Indonesian ground truth using the Indonesian-focused embedding model LazarusNLP/all-indo-e5-small-v4. The average semantic similarity rises from 0.4833 for direct Banyumasan-versus-ground-truth comparison to 0.6427 after translation, producing an absolute gain of 0.1594 and a relative improvement of approximately 33.0% over the baseline. These findings indicate that a structured lexical corpus, when distributed in a directly reusable computational form, can strengthen both resource accessibility and small-scale downstream experimentation for a low-resource regional language.
Keywords: Banyumasan Javanese; Lexical Corpus; Low-resource Language; Python Package; Semantic Similarity;

1. INTRODUCING

Banyumasan Javanese, commonly known as the Ngapak dialect, is a regional language variety that holds significant value as a cultural identity for communities in Banyumas and its surrounding areas. Previous studies have shown that the Ngapak dialect functions not only as a means of communication but also as a representation of Banyumasan cultural identity, including in digital spaces [1] [2]. However, amid the dominance of the Indonesian language and the limited availability of digital language resources for regional languages, Banyumasan faces challenges in terms of documentation and utilization within the digital ecosystem [3]. This condition highlights the increasing importance of providing Banyumasan language resources in digital form [8], [9].

In the field of computational linguistics, the availability of language resources such as corpora and digital lexicons is a fundamental prerequisite for data-driven linguistic analysis, application development, and natural language processing. For low-resource languages,

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

the lack of well-documented resources remains a major obstacle in the development of language technologies [4] [5] [6]. Previous studies on corpus development have also demonstrated that the digitalization of language resources plays a crucial role in documentation, preservation, revitalization, and as a starting point for further data-driven research [7] [8] [9]. Therefore, the development of digital resources for Banyumasan Javanese is relevant not only for linguistic purposes but also for supporting the advancement of regional language technologies.

Although research on Banyumasan has been widely conducted, most studies primarily focus on cultural identity, social functions of language, or its existence in the digital era [1] [2]. Based on the literature review, there is still limited research that explicitly develops structured, validated, and reusable digital lexical resources for Banyumasan that can be directly utilized in computational environments. In contrast, corpus-based research in other languages has progressed toward the development of more systematic resources, including structured data collection, annotation, and reusable documentation [10] [11]. This indicates a gap between the richness of Banyumasan linguistic studies and the availability of ready-to-use digital language resources.

Based on this gap, this study focuses on the development of a structured bilingual lexical resource for Banyumasan Javanese, rather than a large-scale running-text corpus. In this study, the term lexical corpus is used to refer to a curated collection of lexical entries that are organized, normalized, and made reusable for computational processing. However, the resource is more specifically positioned as a Banyumasan-Indonesian bilingual lexicon because each entry consists of a Banyumasan lexical item and its Indonesian equivalent. This differs from a conventional dictionary resource, which is primarily designed for human reference and may include explanations, usage examples, or grammatical information. The present resource transforms dictionary-derived lexical content into a machine-readable dataset that can support lookup, simple translation, and computational experimentation. The dataset consists of 2,000 Banyumasan lexical entries derived from the Banyumasan-Indonesian Dictionary as the primary reference source. Each entry is structured into a formal data format, and its Indonesian equivalent is validated directly based on the dictionary [12]. This approach ensures that the resulting resource has a clear referential basis and maintains a controlled level of linguistic validity [13].

The novelty of this research lies in transforming a dictionary-based lexical resource into a digital resource that is not only presented as a static dataset but also packaged as a Python library that can be installed and reused directly. This approach extends the function of research outputs from mere academic documentation to a practical and reusable resource that can be utilized by researchers and developers [8] [14] . Therefore, this study aims to develop a structured, validated, and computationally accessible digital lexical corpus for Banyumasan Javanese as an initial step toward strengthening digital resources for regional languages.

2. METHODOLOGY

To provide a systematic overview of the research stages, the workflow for constructing the Banyumasan Javanese digital lexical corpus is presented in the form of a flowchart in Figure 1.

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

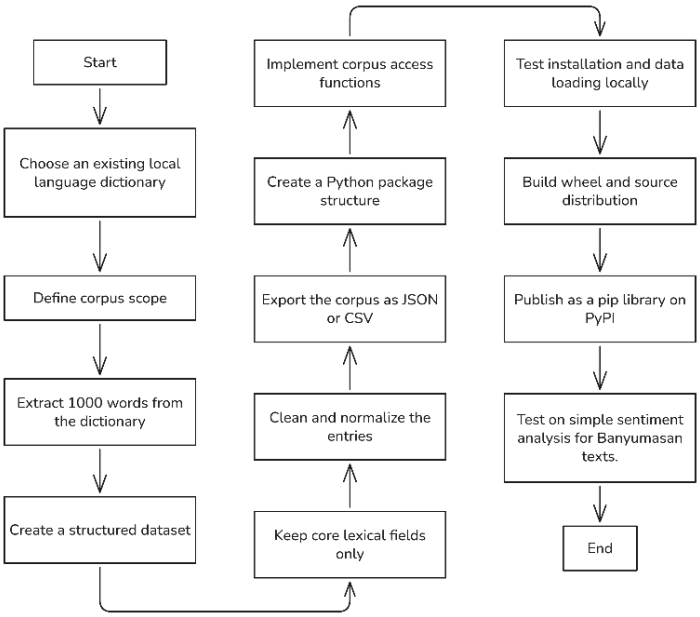

Figure 1. Research Methodology

Based on Figure 1, this study is conducted through three main stages, namely lexical data preparation, digital resource development, and package implementation and testing. These stages are arranged sequentially to ensure a systematic development process, starting from data source selection to the point where the resource can be directly utilized in a Python environment. This approach emphasizes the importance of standardized and reusable data structures in natural language processing [15] [16]

Lexical Data Preparation and Structuring

The study begins with the selection of a regional language dictionary as the primary data source. The dataset is derived from the Banyumasan–Indonesian Dictionary as the main lexical reference. After determining the data source, the scope of the corpus is defined by limiting the dataset to 2,000 Banyumasan lexical entries. This limitation is applied to ensure that the processes of extraction, structuring, and validation can be conducted in a controlled manner, while still producing a sufficiently representative dataset as a digital lexical resource. This stage aligns with corpus development approaches that utilize structured lexical sources as the foundation for dataset construction [17].

The choice to focus on lexical pairs rather than a large running-text corpus reflects the current stage of Banyumasan computational resource development: lexical coverage is a practical starting point because it is easier to validate, easier to normalize, and immediately useful for lookup and simple translation tasks. The canonical curation dataset preserves both runtime fields and provenance-oriented fields. The runtime package later uses a simplified JSON representation for direct loading, while the curation file retains additional metadata useful for maintenance and review. Table 1 summarizes the core fields used during corpus structuring.

Table 1. Dataset Field Definitions

Field	Description
ngapak	Banyumasan lexical item or phrase used as the source-side entry.
indonesia	Indonesian equivalent used as the target-side meaning.
contributor	Contributor identifier retained in the canonical curation file.
source_type	High-level provenance category describing how the entry was obtained.
source_detail	Short traceability note for the lexical source or curation context.
notes	Optional annotation for ambiguity, context, or editorial remarks.

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

Cleaning, Normalization, and Resource Formatting

After extraction, the data were cleaned and normalized into a consistent bilingual format. Incomplete rows were excluded, text was standardized for predictable processing, and the lexical pairs were prepared so they could be validated and exported into a package-friendly representation. The runtime corpus uses a simplified JSON structure that prioritizes direct loading and lookup, while the curation dataset remains richer and easier to maintain.

The resulting packaged corpus contains entry_count = 2000, unique_ngapak = 1996, unique_indonesia = 1444, and duplicate_ngapak_terms = 4. The duplicated Banyumasan headwords are preserved because they reflect legitimate lexical ambiguity in the source material rather than simple noise. This is important for research reuse because low-resource lexical datasets often need to preserve ambiguity explicitly instead of collapsing it away too early. Figure 2 illustrates an example of lexical entries extracted from the dictionary as the primary data source.

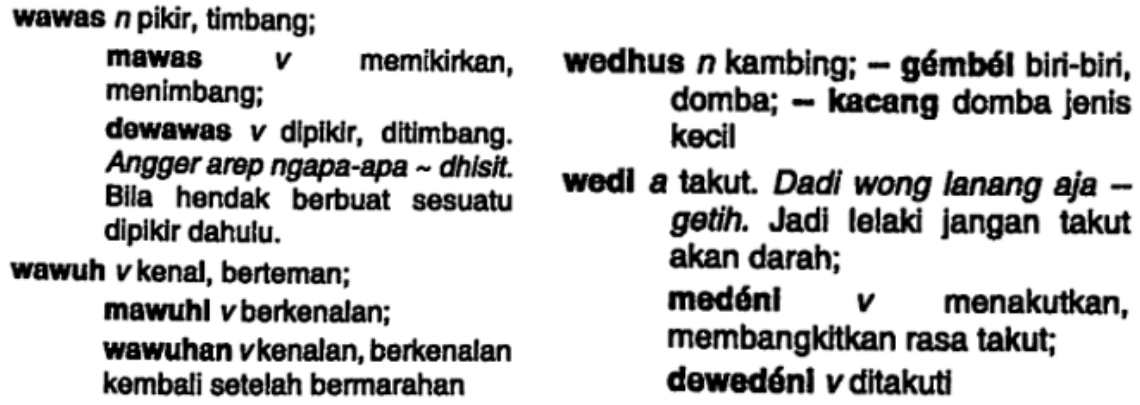
wawas *n* pikir, timbang;
mawas *v* memikirkan, menimbang;
dewawas *v* dipikir, ditimbang. *Angger arep ngapa-apa ~ dhisit.* Bila hendak berbuat sesuatu dipikir dahulu.
wawuh *v* kenal, berteman;
mawuhi *v* berkenalan;
wawuhan *v* kenalan, berkenalan kembali setelah bermarahan
wedhus *n* kambing; -- **gémbél** biri-biri, domba; -- **kacang** domba jenis kecil
wedi *a* takut. *Dadi wong lanang aja -- getih.* Jadi lelaki jangan takut akan darah;
medéni *v* menakutkan, membangkitkan rasa takut;
dewedéni *v* ditakuti

Figure 2. Raw Lexical Data Extracted from the Source Dictionary

Subsequently, data cleaning and normalization processes are performed, including standardizing text formatting by converting all entries into lowercase to ensure structural consistency across the dataset. This step is essential in corpus development to maintain data quality and facilitate integration into NLP systems [18]. The cleaned dataset is then exported into JSON and CSV formats for further processing in the implementation stage.

Table 2. Sample of Normalized Lexical Pairs

Banyumasan	Indonesia
wawas	pikir
wawuh	kenal
wedhus	kambing
wedi	takut

Linguistic Validation Procedure

Linguistic validation was conducted through source-based checking and consistency inspection. Each Banyumasan lexical item and its Indonesian equivalent were checked against the Banyumasan-Indonesian Dictionary to ensure that the dataset followed the reference source. The normalized entries were then inspected to identify incomplete rows, spelling inconsistencies, duplicated entries, and formatting problems. Entries with missing source or target forms were removed, while duplicated Banyumasan headwords were retained when they represented valid lexical ambiguity. This validation procedure supports dictionary-based lexical consistency.

Python Package Development

The curated resource is implemented as the banyumasan-corpus Python package. The package is organized around a small API layer, typed data models, and a bundled JSON corpus, allowing the data to be loaded directly without additional preprocessing. The package is not treated here as a software-engineering centerpiece; rather, it functions as

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

the mechanism that makes the corpus reusable, installable, and reproducible in actual research workflows. In practical terms, the resource can be installed in a standard Python environment with a simple command such as pip install banyumasan-corpus, which lowers the barrier for reuse in later studies.

The implemented research-relevant functions include dataset loading (load_entries), Banyumasan lookup (find_ngapak), Indonesian lookup (find_indonesia), simple translation (translate_ngapak), structured translation analysis (translate_ngapak_detailed), batch translation (translate_ngapak_batch), and corpus statistics (stats). The translation mechanism is deterministic and lexicon-based. Input text is first tokenized using regular expressions, normalized through lowercase conversion and spacing standardization, and then matched against the ngapak field in the packaged JSON corpus. If a token is found, the corresponding Indonesian equivalent is returned. If no match is found, the original token is preserved so that unknown words remain visible instead of being removed or inferred. Ambiguity is handled conservatively: when a Banyumasan entry has more than one possible Indonesian equivalent, the simple translation function uses the available lexical match, while the detailed translation function exposes token-level matching information for manual inspection. Therefore, the package should be understood as a transparent lexical baseline, not a full context-aware translation system [19].

Implementation Example and Evaluation

To demonstrate practical usefulness, the study uses a 100-sentence Banyumasan-Indonesian test set. Each row contains a Banyumasan sentence and an Indonesian ground-truth sentence. The evaluation proceeds in two stages. First, the raw Banyumasan sentence is compared directly with the Indonesian ground truth to establish a baseline semantic similarity score. Second, the Banyumasan sentence is translated using the published banyumasan-corpus package, and the translated output is compared again with the same Indonesian ground truth.

Both comparisons use cosine similarity over sentence embeddings produced by the Indonesian-focused model LazarusNLP/all-indo-e5-small-v4. This evaluation is intentionally modest. It is not a full benchmark over multiple models or translation baselines, but rather a controlled implementation example that tests whether the packaged lexical resource improves semantic alignment in a realistic small-scale usage setting [20].

3. RESULT AND DISCUSSIONS

Lexical Corpus Dataset

The resulting corpus contains 2,000 Banyumasan-Indonesian lexical pairs, 1,996 unique Banyumasan forms, and 1,444 unique Indonesian equivalents. Four Banyumasan headwords remain duplicated in the corpus: dhuwur, dolan, dunya, and enggal. These duplicates were retained because they correspond to genuine ambiguity in meaning or usage rather than accidental duplication. Preserving such ambiguity is useful in lexical resource development because it reflects the structure of the language more faithfully than a forced one-to-one mapping.

The Indonesian side of the corpus also shows meaningful variation. The Indonesian equivalent jatuh appears 16 times, while memukul and ramai each appear 9 times. This indicates that multiple Banyumasan expressions can converge on related Indonesian meanings. For future NLP work, this variation matters because lexical resources must often handle more than one local form for the same or closely related semantic content sample of the dataset is presented in Table 3.

Table 3. Sample of the Lexical Corpus

kata_banyumasan	arti_indonesia
wawas	pikir
dhewek	sendiri
wedhus	kambing

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

wedi	takut
mangan	makan
kencot	Lapar
lunga	Pergi
teka	Datang
nyong	Aku
kepriben	bagaimana

Based on the constructed dataset, it is observed that most entries in the corpus consist of basic vocabulary commonly used in everyday communication, including verbs (e.g., mangan, lunga, teka), nouns (e.g., wedhus), as well as words expressing conditions or emotions (e.g., wedi, kencot). In addition, personal pronouns such as nyong, which are characteristic of Banyumasan, reflect a strong local linguistic identity. This indicates that the developed dataset provides relevant and representative coverage of everyday vocabulary. As an example of dataset utilization, a simple sentence conversion from Banyumasan to Indonesian is performed using a word matching approach. The conversion results are presented in Table 4.

Table 4. Example of Sentence Conversion Results

Banyumasan Sentence	Converted Result
nyong kencot arep mangan	aku lapar ingin makan
dhewek lunga teka omah	sendiri

Table 4 demonstrates that the lexical corpus is capable of directly mapping Banyumasan words into Indonesian equivalents. However, the resulting translations are still literal and do not fully account for grammatical structure, leading to sentences that may not completely conform to standard Indonesian syntax. Nevertheless, the constructed lexical corpus serves not only as a dataset but also as a resource with potential applications in dictionary-based translation, basic text analysis, and the development of Natural Language Processing (NLP) systems for Banyumasan.Further analysis reveals that a single Indonesian meaning can be represented by multiple Banyumasan lexical variants. For instance, the word "jatuh" appears 16 times in the dataset, while "memukul" and "ramai" each appear 9 times. This finding indicates that Banyumasan exhibits a relatively high degree of lexical variation in expressing similar meanings.

Such variation is an important aspect in the development of NLP systems, as models need to recognize not only a single word form but also multiple lexical variants with similar meanings. With a structured lexical corpus, systems can perform more accurate word-to-meaning mapping, thereby improving semantic understanding in Banyumasan texts.
Therefore, the developed corpus does not merely function as a digital dictionary, but also serves as an initial foundation for enhancing semantic representation in computational processing of the Banyumasan language.

Python Package Development

In this study, Packaging the corpus as banyumasan-corpus changes the resource from a static research output into a directly reusable artifact. Researchers can load the entries programmatically, inspect lexical mappings, retrieve Banyumasan forms by Indonesian meaning, and perform simple lexicon-based translation without rebuilding the dataset pipeline. This matters because reproducibility in low-resource language work often fails at the resource-distribution stage: a dataset may be described in a paper but still be difficult to use in practice.

The final result indicates that the package has been successfully built in the form of wheel and source distribution, allowing it to be installed and reused. Thus, the Banyumasan lexical corpus has evolved from a static dataset into a practical digital resource for natural language processing applications. The main functions provided by the package are summarized in Table 5.

Table 5. Main Functions of the Lexical Corpus Python Package

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

Function	Description
load_entries()	Loads the entire lexical corpus dataset
find_ngapak()	Retrieves the Indonesian meaning of a Banyumasan word
find_indonesia()	Finds Banyumasan lemmas based on Indonesian meanings
translate_ngapak()	Performs simple translation of Banyumasan text into Indonesian
translate_ngapak_batch()	Translates multiple Banyumasan texts simultaneously
translate_ngapak_detailed()	Provides detailed translation results, including token matching
stats()	Displays statistical information about the dataset

Implementations and Usage of the Package

The developed package is implemented to perform simple text processing for Banyumasan Javanese by utilizing corpus-based lookup and translation functions.

The find_ngapak function is used to retrieve the Indonesian meaning of Banyumasan words directly from the corpus, while find_indonesia enables searching Banyumasan lemmas based on Indonesian meanings. In addition, simple translation is performed using the translate_ngapak function with a token-based matching approach, where the input text is tokenized using regular expressions, normalized, and matched with corpus entries. An example of the function usage is presented in Figure 3.

```
#Input
from banyumasan_corpus import find_ngapak, translate_ngapak

# mencari arti kata
hasil = find_ngapak("wawas")
print(hasil)

# translasi kalimat sederhana
kalimat = "nyong kencot arep mangan"
terjemahan = translate_ngapak(kalimat)
print(terjemahan)

#Output
pikir
aku lapar ingin makan
```

Figure 3. Example Usage of the Lexical Corpus Python Package

In practice, the package can already support small experiments beyond simple lookup. For example, translating the sentence Woh pelem neng wit kuwi wis mateng kabeh yields Buah mangga di pohon kuwi sudah matang semua, which preserves most of the core meaning while still leaving one local token untranslated. This untranslated token illustrates the unknown-word handling mechanism in the package, where unmatched tokens are preserved in the output to make lexical coverage gaps visible for future corpus expansion. The more detailed translation output also exposes chunk-level alignment, making it easier to inspect which parts of the sentence were resolved by the corpus and which parts still require further lexical coverage or disambiguation.

Open Resource and Community Extension

An additional strength of the resource is that it is designed to be openly reusable and extendable. Because the corpus is distributed as an installable Python package, it is publicly accessible through the Python Package Index (PyPI) at https://pypi.org/project/banyumasan-corpus/, allowing other researchers, students, and developers can adopt it directly in their own experiments and then improve it through broader lexical coverage, better ambiguity handling, richer annotation, or more demanding

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

evaluation settings. This open-resource orientation is especially important for low-resource languages, where progress often depends on incremental community contribution rather than on one isolated dataset release.

Semantic similarity Evaluation

Table 6 contains a small handcrafted Banyumasan test case that mirrors the structure used in the larger evaluation: a source sentence, an Indonesian reference, and the translation generated by the package. The examples show three typical situations: a mostly successful lexical transfer, a partially translated sentence with unresolved pronouns, and a difficult case where important tokens remain untranslated.

Table 6. Sample Results of Translation Evaluation

Banyumasan input	Indonesian reference	Example Package Translation
Woh pelem wis mateng	Buah mangga sudah matang	Buah mangga sudah matang
Weteng enyong kencot	Perut saya lapar sekali	Perut enyong lapar
Ngesuk nyong tampa rapot	Besok saya tidak ada rapat	Besok nyong tampa rapot
Woh pelem wis mateng	Buah mangga sudah matang	Buah mangga sudah matang

Table 7 shows the full 100-sentence implementation example provides an initial quantitative view of how useful the packaged corpus is for translation-oriented processing. When the original Banyumasan rows are compared directly against Indonesian ground truth, the average semantic similarity is 0.4833. After translation with banyumasan-corpus, the average semantic similarity increases to 0.6427. This corresponds to an absolute gain of 0.1594 and a relative improvement of approximately 33.0% over the baseline.

Table 7. Semantic Similarity Evaluation Results

Metric	Value
Rows Evaluated	100
Embedding model	LazarusNLP/all-indo-e5-small-v4
Baseline similarity	0.48333
Post-translation similarity	0.6427
Absolute improvement	0.1594
Relative Improvement over baseline	33.0%

The row-level results show that the gain is meaningful but not uniform. One strong positive example is Woh pelem neng wit kuwi wis mateng kabeh, where similarity rises from 0.1277 to 0.8105 after translation. A more moderate example is Ngesuk nyong tampa rapot, where the gain is from 0.1637 to 0.3251. However, there are also difficult cases. For "Oh kaya kue, ya sing sabar ya mbekayu" , the score decreases from 0.6812 to 0.5240 because literal lexical substitution and unresolved context still distort the final sentence meaning. These mixed outcomes are consistent with the design of the system: it is a lexical baseline built for accessibility and reuse, not a fully context-aware translation model.

The evaluation should therefore be read in moderate terms. It demonstrates that the packaged corpus improves semantic alignment on average in a practical 100-sentence example, but it does not yet establish full sentence-level translation quality or superiority over other approaches. Even so, the improvement is large enough to show that the resource is already useful as a baseline computational artifact and as a starting point for future Banyumasan NLP research.

Study Limitations

This study has several limitations. First, the corpus is limited to 2,000 lexical entries, so it does not yet represent the full vocabulary coverage of Banyumasan Javanese. Second, the resource preserves lexical ambiguity but does not yet resolve it contextually, which may cause literal or incomplete translations. Third, the evaluation uses 100 sentence

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

examples and one Indonesian-focused embedding model, so the results should be read as an initial implementation test rather than a comprehensive translation benchmark. Future evaluation should include larger test sets, human judgment, comparison with alternative baselines, and more detailed error analysis.

4. CONCLUSION

This study developed a structured Banyumasan-Indonesian lexical corpus with 2,000 entries and implemented it as the reusable Python package banyumasan-corpus. The resource supports dataset loading, bilingual lookup, simple translation, translation analysis, batch processing, and corpus statistics, making it directly usable for computational experiments instead of remaining only as a descriptive dataset.

The implementation example on 100 sentences shows that translation with the packaged corpus improves average semantic similarity from 0.4833 to 0.6427, with an absolute gain of 0.1594 and a relative improvement of about 33.0%. These results do not imply that the current package solves Banyumasan-to-Indonesian translation comprehensively, but they do show that a structured lexical corpus can deliver measurable practical value when made computationally reusable.

Future work should extend the resource through broader lexical and phrasal coverage, expert or native-speaker validation, and larger evaluation settings. Further development may also include sentence-level translation refinement, richer annotation, human evaluation, comparison with alternative baselines, and broader downstream tasks such as retrieval, classification, and low-resource lexical induction.

5. REFERENCES

[1] V. I. F. Maulidha and H. B. Thontowi, "The Influence of Familial Ethnic Socialization on Self-Esteem among Banyumasan Javanese Adolescents as Mediated by Ethnic Identity," *Jurnal Psikologi*, vol. 52, no. 3, p. 295, 2025, doi: 10.22146/jpsi.109320.

[2] C. Nugroho and I. P. Kusuma, "Identitas Budaya Banyumasan dalam Dialek Ngapak," *Jurnal Ilmu Komunikasi*, vol. 21, no. 2, pp. 333–347, Sep. 2023, doi: 10.31315/JIK.V21I2.4556.

[3] A. G. Pawestri, "Membangun Identitas Budaya Banyumasan Melalui Dialek Ngapak Di Media Sosial," *Jurnal Pendidikan Bahasa dan Sastra*, vol. 19, no. 2, pp. 255–266, May 2020, doi: 10.17509/bs_jpbsp.v19i2.24791.

[4] E. N. Purba, D. P. Togatorop, and A. Rosari, "Analisis Keterbatasan Korpus Bahasa dan Pemanfaatan AI dalam Penyusunan Kamus Batak Toba," *Jurnal Pendidikan Tambusai*, vol. 9, no. 3, pp. 36479–36490, 2025.

[5] A. Rokhman, R. E. Priyono, I. Santosa, S. Pangestuti, & Mustasyfa, and T. Kariadi, "Existence of Banyumasan Javanese Language in Digital Era," *Humanities and Social Science Research*, vol. 5, no. 2, pp. p1–p1, May 2022, doi: 10.30560/HSSR.V5N2P1.

[6] S. Cahyawijaya *et al.*, "NusaCrowd: Open Source Initiative for Indonesian NLP Resources," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 13745–13818, 2023, doi: 10.18653/V1/2023.FINDINGS-ACL.868.

[7] H. Handoko, "Developing the Corpus of Minangkabau Language: Insights, Challenges, and Future Directions," *Jurnal Arbitrer*, vol. 11, no. 3, pp. 413–429, Sep. 2024, doi: 10.25077/AR.11.3.413-429.2024.

[8] B. K. Bal, B. Prasain, R. R. Ghimire, and P. Acharya, "Strategies for Corpus Development for Low-Resource Languages: Insights from Nepal," *Automatic Speech Recognition and Translation for Low Resource Languages*, pp. 297–330, Jan. 2024, doi: 10.1002/9781394214624.CH15.

[9] A. E. Nanda, V. P. Rantung, and K. Santa, "Development of a Web-Based Batak Simalungun Regional Language Corpus Using the Rapid Application Development

Nisrina Hanifa Setiono: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 160-169
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1705

Method," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 549–558, 2024, doi: 10.52436/1.jutif.2024.5.4.2210.

[10] D. I. Adelani *et al.*, "MasakhaNER: Named Entity Recognition for African Languages," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 1116–1131, Oct. 2021, doi: 10.1162/TACL_A_00416.

[11] C. Pramartha and others, "Preserving the Enggano Language: A Digital Dictionary Approach," *ACIS Proceedings*, pp. 1–8, 2024.

[12] A. Tohari, P. Suratno, and A. Sudono, *Kamus Bahasa Jawa Banyumasan Indonesia*, Cetakan pe. Semarang: Balai Bahasa Provinsi Jawa Tengah, 2014.

[13] S. Conggresco, V. P. Rantung, and Q. C. Kainde, "Development of a Web-Based Tonsea Language Corpus Using the Evolutionary Prototyping Method," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 535–542, 2024, doi: 10.52436/1.jutif.2024.5.4.2201.

[14] M. W. Goodman and F. Bond, "Intrinsically interlingual: The Wn Python library for wordnets," *GWC 2021 - Proceedings of the 11th Global Wordnet Conference*, pp. 100–107, 2021, doi: 10.18653/v1/2021.gwc-1.12.

[15] G. I. Winata *et al.*, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 815–834, May 2022, doi: 10.18653/v1/2023.eacl-main.57.

[16] J. M. List, R. Forkel, S. J. Greenhill, C. Rzymski, J. Englisch, and R. D. Gray, "Lexibank, a public repository of standardized wordlists with computed phonological and lexical features," *Scientific Data 2022 9:1*, vol. 9, no. 1, pp. 316-, Jun. 2022, doi: 10.1038/s41597-022-01432-0.

[17] Y. Prabowo, M. Gabriel, Nazarudin, T. Ratumanan, and M. Maslim, "Preserving Meher and Woirata Corpus Languages using Neural Machine Translation," *Indonesian Journal of Information Systems*, vol. 6, no. 2, pp. 156–161, Feb. 2024, doi: 10.24002/IJIS.V6I2.8542.

[18] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Word Level Language Identification in Indonesian-Javanese-English Code-Mixed Text," *Procedia Comput. Sci.*, vol. 244, pp. 105–112, Jan. 2024, doi: 10.1016/J.PROCS.2024.10.183.

[19] K. Resiandi, Y. Murakami, and A. H. Nasution, "Neural Network-Based Bilingual Lexicon Induction for Indonesian Ethnic Languages," *Applied Sciences 2023, Vol. 13, Page 8666*, vol. 13, no. 15, p. 8666, Jul. 2023, doi: 10.3390/APP13158666.

[20] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," Feb. 2024, Accessed: Apr. 16, 2026. [Online]. Available: https://arxiv.org/pdf/2402.05672

Nisrina Hanifa Setiono: *Corresponding Author