ISSN 2664-4150 (Print) & ISSN 2664-794X (Online)
**South Asian Research Journal of Engineering and Technology**
*Abbreviated Key Title: South Asian Res J Eng Tech*

**| Volume-6 | Issue-1 | Jan-Feb- 2024 |** **DOI: 10.36346/sarjet.2024.v06i01.001**

**Original Research Article**

# Using Big Data Approach to Create a Treebank of Informal and Formal Indonesian

Danang Satria Nugraha[1*]
[1]Sanata Dharma University, Indonesia & University of Szeged, Hungary

***Corresponding Author:** Danang Satria Nugraha
Sanata Dharma University, Indonesia & University of Szeged, Hungary

**Article History**
Received: 29.11.2023
Accepted: 03.01.2024
Published: 04.01.2024

**Abstract:** This research presents a comprehensive Big Data Approach that was utilized to create a Treebank of Informal and Formal Indonesian (TINTA). The study focuses on the dynamic spectrum of language usage in Indonesia. It incorporates extensive data collection, preprocessing, and annotation strategies to construct a dual-tiered corpus encompassing formal and informal linguistic expressions. Through leveraging advanced computational techniques, the creation of TINTA aims to capture the nuanced variations in Indonesian language structures across diverse contexts. This annotated treebank provides a valuable resource for advancing Natural Language Processing (NLP) applications and linguistic research endeavors by facilitating more profound insights into the grammatical intricacies and semantic nuances prevalent in informal and formal Indonesian language.

**Keywords:** Big data, Corpus development, Formal and informal language, Indonesian language, Linguistic annotation.

## INTRODUCTION

Language is a complex and ever-changing system influenced by various cultural, social, and contextual factors. In Indonesia, a diverse range of formal and informal language usage offers an excellent opportunity for exploration (Nugraha, 2020a, 2021b, 2021a). We are creating a comprehensive Treebank of Informal and Formal Indonesian, or TINTA, to capture this complex linguistic landscape using extensive data methodologies. This project requires a nuanced approach to capture the intricate variations in language, from structured formal communication to the fluidity of informal discourse.

Indonesia boasts a rich linguistic diversity encompassing a wide range of geographical, socio-cultural, and contextual domains. However, this diversity poses unique challenges for computational linguistics (Nugraha, 2020b, 2020c, 2023). The absence of a comprehensive resource that captures the formal and informal aspects of the Indonesian language has been a significant hurdle in developing adequate Natural Language Processing (NLP) tools and models tailored to this intricate linguistic ecosystem. TINTA, a dual-tiered corpus that takes advantage of extensive data collection, annotation, and computational techniques to reflect the multifaceted nature of Indonesian communication, has been developed to address this challenge. This initiative will pave the way for developing more effective NLP tools and models better suited for the Indonesian language ecosystem.

The article discusses the creation process of TINTA, a treebank annotated using big data techniques to capture the diversity of language patterns. This resource is invaluable for examining formal and informal Indonesian languages' grammatical structures, semantic nuances, and pragmatic usage. By incorporating a vast amount of data, TINTA can provide a comprehensive view of the language, making it an ideal resource for further research in linguistics and natural language processing applications. The use of big data techniques in creating TINTA has enabled researchers to explore and analyze linguistic phenomena in ways that were not previously possible, thus opening up new avenues for understanding the complexities of the Indonesian language.

# LITERATURE REVIEW

**Theoretical Framework**

Remarkable progress in computational linguistics can be attributed to developing linguistic resources like language-specific treebanks. These resources have enabled the creation of annotated corpora and treebanks for several languages, leading to advancements in Natural Language Processing (NLP) applications (Claridge, 2007; Hoffmann, 2007; McEnery, 2000; Renouf *et al*., 2007; Rosenbach, 2007). However, the Indonesian language's complex variations between formal and informal language make it a unique yet underexplored domain in this context. The development of treebanks has primarily focused on languages with well-defined grammatical structures, often neglecting the nuances present in informal communication (Fletcher, 2007; Hundt *et al.,* 2007; Lüdeling *et al*., 2007). This disparity presents specific difficulties in leveraging the full capabilities of computational models for languages like Indonesian, where the interplay between formal and informal registers significantly impacts communication patterns.

Recent research has emphasized the need for extensive linguistic resources covering formal and informal language variations. Studies on developing annotated corpora for informal language in various linguistic contexts have demonstrated the importance of capturing colloquial expressions, pragmatic nuances, and diverse syntactic structures inherent in informal communication (Mizumoto *et al*., 2020; Römer & O'Donnell, 2011; Rühlemann & O'Donnell, 2012). Such efforts aim to bridge the gap in language resources and address the challenges in natural language processing and machine learning for informal text (Baayen & Linke, 2020; Gries & Paquot, 2020; Levshina, 2020).

The field of big data has brought about significant changes in how we collect, process, and analyze vast amounts of linguistic data (Holmes, 2017b, 2017e, 2017c). By utilizing advanced techniques in corpus development, we can now create more comprehensive representations of language diversity, resulting in more robust linguistic resources for computational analysis and NLP applications (Bühlmann *et al*., 2016; Holmes, 2017g).

The field of computational linguistics and corpus development has been proliferating (Ädel, 2020; Jaworska, 2016; McEnery & Hardie, 2011; Newman & Cox, 2020; Zeldes, 2020). However, the Indonesian language ecosystem needs more attention to comprehensive resources that include formal and informal language expressions. This literature review highlights the urgent need and potential benefits of a Big Data Approach in building a Treebank of Informal and Formal Indonesian (TINTA) to bridge this gap. TINTA would enable significant progress in Indonesian language processing and linguistic research.

**Previous Research**

Existing studies on corpus development for the Indonesian language have primarily concentrated on either the formal or informal language registers in isolation, resulting in a fragmented portrayal of the language landscape (Baig *et al*., 2020; Chambers, 2019; Franzosi, 2021; Peter *et al*., 2023). This lack of comprehensive resources integrating formal and informal expressions impedes the progress of Natural Language Processing (NLP) applications customized to this diverse language ecosystem.

Numerous endeavors have been undertaken to curate formal language corpora for Indonesian. The primary focus of these initiatives has been to comprehend the structured patterns of language. However, these efforts need to recognize the lively and widespread informal communication styles deeply ingrained in Indonesian society. Such communication styles can offer valuable insights and can be pivotal to developing a more comprehensive understanding of the Indonesian language. Studies on informal language in Indonesia have yielded significant findings on colloquial expressions, idiomatic phrases, and the pragmatic use of language. However, these studies are typically constrained in scope and scale and must be adequately annotated or integrated into broader linguistic resources. As such, there is a need for comprehensive and integrated research in this area to advance our understanding of informal language use in Indonesia.

The development of treebanks for different languages worldwide has significantly contributed to the field of natural language processing. However, more attention needs to be given to the duality of language usage (formal and informal) within a single corpus, with only a few projects addressing this issue. In corpus linguistics, it has become increasingly essential to incorporate informal language nuances in addition to formal structures. This is supported by studies in various linguistic contexts, including English, Spanish, and Mandarin, demonstrating the necessity of annotated resources encompassing formal and informal language. Such resources are essential for developing robust natural language processing (NLP) applications and carrying out in-depth linguistic analysis.

The lack of a unified and comprehensive treebank that incorporates both formal and informal Indonesian language is a significant obstacle in developing effective computational models and tools that can handle the intricacies of Indonesian communication. Despite ongoing efforts, this challenge continues to impede progress in the field. Our research aims to use a Big Data Approach for constructing a Treebank of Informal and Formal Indonesian (TINTA), which would help develop

Indonesian language processing and linguistic studies. This would help bridge the existing gap in the field of Indonesian language processing.

# METHODOLOGY

### Data Collection and Compilation

Our primary objective is to acquire Big Data for the Indonesian language. To achieve this, we utilize diverse sources, including social media platforms, online forums, news articles, and written publications, to collect many Indonesian language samples. Once we have collected the data, we employ advanced machine learning algorithms and linguistic markers to segregate the data into formal and informal text segments. This process involves identifying and distinguishing between various language expressions that are commonly used in formal and informal settings. Lastly, the segregation of formal and informal language expressions is essential in ensuring the data is correctly analyzed and interpreted. This helps us identify patterns and trends functional in various fields, such as marketing, customer service, and social science research.

### Data Preprocessing

Normalization and Cleaning involves removing irrelevant, extraneous, or noisy data from the raw data, correcting spelling errors and standardizing the data format to ensure consistency and quality. On the other hand, Tokenization and Annotation involve segmenting textual data into individual tokens and then annotating them for various linguistic features, such as parts of speech, syntactic structures, and semantic elements, using established linguistic frameworks and tools.

### Dual Annotation Framework

Formal Annotation involves the application of grammatical rules, syntactic structures, and other formal linguistic conventions to annotate the corpus's formal language subset. On the other hand, Informal Annotation involves incorporating colloquialisms, idiomatic expressions, and pragmatic elements into the annotation process to match the characteristics of informal communication styles.

### Integration and Alignment

The task at hand is to merge formal and informal language annotations to preserve their distinctive features and unify them into a single corpus. This can be achieved by developing effective integration strategies that align the annotated subsets while maintaining the differences between formal and informal language segments. To ensure the accuracy and consistency of the merged corpus, rigorous cross-validation and consistency checks need to be performed. These validation processes help establish the coherence and alignment of formal and informal annotations, ensuring that the final corpus is high quality and suitable for various applications.

### Corpus Expansion and Iterative Refinement

Incremental Data Addition involves adding new data to the existing corpus, which helps expand the dataset and improve its representativeness towards the latest language usage trends. Annotation Quality Assurance (AQA) uses iterative refinement techniques to ensure the highest annotation accuracy and consistency.

### Treebank Construction

Treebank Construction generates hierarchical structures, syntactic trees, and dependency graphs for annotated formal and informal language segments. This is done by incorporating linguistic features into the treebank representations to capture nuances and variations in Indonesian language usage. It requires a deep understanding of linguistics and natural language processing techniques to represent the complex structures and features of the language accurately.

### Evaluation and Validation

The text describes two essential measures used in natural language processing. The first measure, Intra-Annotation Agreement, assesses the level of agreement between multiple annotators when performing formal and informal annotation tasks. The second measure, Inter-Annotation Consistency, evaluates the degree to which formal and informal annotations within a corpus align with each other and form a coherent whole. These measures are crucial for ensuring the accuracy and reliability of natural language processing techniques.

# RESULTS AND DISCUSSION

### Effectiveness of Big Data Approach in Distinguishing and Integrating Language Expressions

The Big Data Approach accurately differentiated between formal and informal Indonesian language expressions and created separate subsets within the TINTA corpus. Moreover, the integration process of formal and informal annotations was exact and maintained the distinction between both language registers within the unified corpus (O'Keeffe & McCarthy, 2022; Paquot & Gries, 2020; Th. Gries, 2020). This showcases the effectiveness and reliability of the Big Data Approach in processing and merging large linguistic datasets with high accuracy. By utilizing a Big Data Approach,

which incorporates various sources such as social media, online forums, and written publications, we efficiently differentiated between formal and informal Indonesian language expressions. Using machine learning algorithms and linguistic markers, we could effectively distinguish between the two registers, which facilitated the formation of distinct subsets within the TINTA corpus.

Integrating formal and informal annotations in a unified corpus has shown significant accuracy. Computational methods were used to merge annotated subsets while maintaining the inherent characteristics of each register. Special techniques were employed to maintain segregation during the integration phase, successfully preserving the distinct nature of formal and informal language expressions within TINTA. The integrated treebank produced a reliable and consistent output that preserved the accuracy and completeness of both formal and informal language segments. Even after integration, the formal and informal annotations were still distinguishable, creating a comprehensive resource for researchers and NLP practitioners to access the full range of Indonesian language variations.

The successful implementation of the Big Data Approach has enabled the efficient differentiation, labeling, and combination of formal and informal Indonesian language expressions within TINTA. This is a significant advancement in developing a consolidated resource for Indonesian language processing and linguistic analysis. The successful application of the Big Data Approach in distinguishing formal and informal Indonesian language expressions within TINTA has paved the way for a significant advancement in corpus development methodologies. By leveraging diverse data sources, including social media, online forums, and written publications, this approach allowed for a comprehensive representation of language variations prevalent in Indonesian communication.

The Big Data Approach enabled a nuanced segmentation of language registers, creating distinct subsets within TINTA. Machine learning algorithms and linguistic markers effectively identified and segregated formal and informal language expressions, ensuring the preservation of their inherent characteristics. The integration process, guided by computational techniques, maintained the integrity of both registers while harmoniously merging formal and informal annotations into a unified corpus (Anthony, 2020; Gries & Durrant, 2020). The scalability and adaptability of the Big Data Approach empowered TINTA's construction to encompass a broad spectrum of linguistic diversity. Using computational methods ensured the corpus's representativeness of contemporary language trends, making it a dynamic resource reflecting the evolving nature of Indonesian communication.

This successful integration of formal and informal language expressions within TINTA shows the feasibility and efficacy of employing a Big Data Approach in corpus development. The corpus is a testament to the potential of leveraging diverse data sources and computational techniques, laying the groundwork for more accurate and comprehensive linguistic resources tailored to intricate language variations. As such, it sets a precedent for future endeavors in corpus development across diverse linguistic contexts and languages, showcasing the power of extensive data methodologies in capturing the nuances of language.

**Representation of Linguistic Features in Informal Indonesian Language**

The TINTA treebank, designed to capture the informal Indonesian language, demonstrated excellent proficiency in encapsulating a wide range of colloquial expressions, idiomatic phrases, and pragmatic elements unique to informal communication. The corpus included diverse syntactic structures, non-standard grammar, and informal lexical choices reflective of everyday conversational Indonesian language, significantly portraying informal linguistic features. Its comprehensive coverage of idiomatic phrases, pragmatic nuances, and contextual language usages specific to informal communication provided a nuanced understanding of the pragmatic aspects of the informal Indonesian language. Overall, this repository of linguistic features in the informal Indonesian language showcases TINTA's ability to capture the intricate nuances of informal communication, empowering advanced NLP applications and linguistic analyses within this domain.

The TINTA treebank has demonstrated an impressive ability to capture many colloquial expressions commonly used in informal Indonesian conversations. The corpus includes a broad spectrum of informal linguistic features, from slang to regional vernacular. As a result, the annotated segments are enriched with a diverse representation of colloquialisms. TINTA, a corpus of the informal Indonesian language, successfully represented a wide range of syntactic structures characteristic of informal communication styles. The corpus included various sentence constructions, non-standard grammatical forms, and informal lexical choices, all commonly used in everyday conversations. By incorporating such linguistic features, TINTA significantly contributed to portraying informal language use in Indonesian, making it an invaluable resource for researchers studying Indonesian linguistic patterns.

TINTA is a corpus that represents the informal Indonesian language, and it has successfully captured a broad range of syntactic structures characteristic of informal communication styles. The corpus includes various sentence constructions, non-standard grammatical forms, and informal lexical choices commonly used in everyday conversations. By incorporating such linguistic features, TINTA has significantly contributed to portraying informal language use in

Indonesian. As a result, it has become an invaluable resource for researchers studying Indonesian linguistic patterns. The advanced treebank of the informal Indonesian language developed by TINTA accurately represents various linguistic features, including colloquial expressions, idiomatic phrases, syntactic variations, and pragmatic elements. This comprehensive coverage enhances the Indonesian language resources, providing a solid foundation for advanced NLP applications and linguistic analyses in this domain.

Integrating colloquial descriptors, syntactic variances, idiomatic expressions, and pragmatic components in TINTA's informal Indonesian language segments establishes a significant milestone in corpus development. This amalgamation enriches the depth and authenticity of the corpus by providing a nuanced portrayal of informal communication styles that are widespread in Indonesian society (Gut, 2020; Lefer, 2020). TINTA is a corpus that extensively represents the linguistic features unique to the informal Indonesian language. The corpus is characterized by its depth and authenticity, as it captures a diverse range of colloquial expressions, syntactic variations, idiomatic phrases, and pragmatic elements. This comprehensive coverage offers a nuanced portrayal of the complex communication styles prevalent in Indonesian society.

Incorporating colloquial language and informal vocabulary in the corpus enhances its genuineness, mirroring the organic communication patterns observed in routine social interactions. The all-inclusive nature of TINTA makes it a valuable asset for both computational linguistics and sociolinguistic research, as it provides a window into the socio-cultural nuances embedded within informal language use. TINTA's corpus is a valuable resource for studying the varied syntactic structures found in Indonesian language variations, particularly in informal communication. It comprehensively represents non-standard grammar, different sentence constructions, and regional vernacular, providing a detailed understanding of the syntactic complexities inherent in informal discourse.

Through its incorporation of idiomatic phrases and pragmatic elements, TINTA significantly enhances its utility for pragmatic analysis and language understanding. By cataloging context-specific language usages and capturing the pragmatic nuances embedded in informal communication, the corpus provides a valuable resource for exploring the intricacies of language beyond its structural aspects. Its ability to identify and analyze contextual factors and their impacts on language usage makes it a valuable tool for researchers and experts looking to delve deeper into the complexities of language. The TINTA corpus strongly represents linguistic features unique to the informal Indonesian language. This makes it an essential resource for researchers in computational linguistics, sociolinguistics, and pragmatic analysis within the Indonesian linguistic context. Its comprehensive and authentic portrayal of these features positions it as a foundational tool for further research.

**Alignment and Coherence between Formal and Informal Annotations**

The TINTA corpus exhibits a remarkable consistency in its formal and informal annotations, showcasing an impressive alignment and coherence between the two registers. The cross-validation results confirm the accuracy and reliability of the formal and informal annotation tasks, as evidenced by the strong agreement scores observed during intra-annotation and inter-annotation evaluations (Hirschheim & Whitchurch, 2023; Holmes, 2017a, 2017d). TINTA exhibited remarkable consistency and alignment between formal and informal annotations throughout the corpus. Formal annotations correctly represented structured language patterns, while informal annotations captured colloquial Indonesian nuances and diverse expressions. Despite the contrasting attributes, the annotations were coherent and aligned, resulting in an equitable representation of both language registers (Holmes, 2017f; Shalabh, 2019; Xodabande *et al*., 2023).

The evaluations among annotators demonstrated high levels of agreement in annotating both formal and informal language segments. This agreement indicates a consistent and standardized approach to annotation, which enhances the reliability and accuracy of the annotated corpus. Such consistency minimizes discrepancies and strengthens the robustness of the linguistic resource (Kehoe, 2020; Knight & Adolphs, 2020). The evaluation comparing formal and informal annotations showed that TINTA has effectively integrated both subsets without compromising their differentiation. This reinforces the authenticity and versatility of the corpus, which can be utilized for various natural language processing applications and linguistic analyses.

The TINTA corpus has high accuracy and reliability due to the careful alignment of its formal and informal annotations. It comprehensively represents the nuances of Indonesian language variations, with a consistent and balanced coverage of both registers. This makes it an ideal resource for conducting in-depth research and developing advanced NLP models that are contextually sensitive and capable of accurately processing the Indonesian language. The TINTA corpus is a dependable and practical resource for linguistic research and natural language processing (NLP) applications, thanks to the careful alignment and coherence between formal and informal annotations. The accuracy and credibility of the corpus are supported by the consistency in intra-annotation agreement and inter-annotation evaluations, which ensure that the corpus can be harmoniously integrated while preserving the distinctiveness of formal and informal language expressions. As a result, TINTA is a reliable and unified corpus that represents a wide range of variations in the Indonesian language.

The high level of consistency observed in intra-annotation agreement and inter-annotation evaluations is a testament to the accuracy and durability of the corpus. The strong agreement scores between annotators for formal and informal language segments validate the standardized approach employed in annotating TINTA, guaranteeing a consistent representation of linguistic features throughout the corpus. Integrating formal and informal annotations in the Indonesian language corpus is highly aligned and coherent, ensuring consistency and balance. The corpus maintains a clear distinction between formal and informal language expressions, providing researchers and NLP practitioners with an accurate resource that captures the diverse nuances of Indonesian communication styles. Through its effective integration, TINTA balances the need for consistency and distinctiveness, making it a suitable platform for various linguistic studies and NLP applications. Researchers aiming to investigate language variations or build contextually sensitive computational models can leverage TINTA's consistent performance to obtain accurate insights and develop sturdy language processing systems.

TINTA corpus exhibits a high level of alignment and coherence between its formal and informal language annotations, which creates opportunities for cross-disciplinary research. This feature enables collaborations between computational linguists, sociolinguists, and discourse analysts. The nuanced representation of formal and informal language expressions offers a promising platform for investigating language phenomena in various socio-cultural contexts. The meticulous integration of distinctions preserves the credibility of TINTA as a unified and reliable corpus, opening avenues for comprehensive linguistic research and facilitating the development of advanced NLP models tailored to the intricacies of Indonesian language variations.

## CONCLUSION

Creating the Treebank of Informal and Formal Indonesian (TINTA) is a significant achievement in developing Indonesian language corpora. This project utilized a Big Data Approach to construct a comprehensive resource that successfully segregates, integrates, and balances formal and informal language expressions. TINTA has immense potential for advancing Indonesian language processing and linguistic studies due to its distinguished features and thoroughness.

It is important to acknowledge certain limitations of TINTA. Although the corpus has a broad scope, it may still have some underrepresentation, especially when capturing the ever-evolving nature of informal language. Additionally, potential biases in data collection sources and the challenges in accurately annotating subtle pragmatic nuances may impact the completeness of the corpus. Despite efforts to maintain coherence, there may still be areas where further refinements are needed in aligning formal and informal annotations.

To enhance the utility of TINTA and overcome its limitations, future research can be pursued in several aspects. Firstly, additional data sources can be incorporated, and annotations can be refined to encompass a broader range of language variations and sociocultural contexts. Secondly, multi-layered annotations can be introduced to provide a more comprehensive understanding of Indonesian language usage, including deeper linguistic aspects such as sentiment analysis, discourse structure, or pragmatic elements. Thirdly, context-specific analyses can be conducted to explore the influence of regional dialects, sociolinguistic factors, and domain-specific language usage on formal and informal Indonesian communication. Finally, benchmarking studies and evaluations can be initiated to assess the applicability and effectiveness of TINTA in diverse NLP tasks and linguistic analyses.

In summary, TINTA serves as a remarkable example of the power of utilizing a Big Data methodology to create a linguistically diverse and well-balanced corpus. While it has limitations, this resource is an essential tool for progressing Indonesian language research and facilitating the creation of more precise and contextually aware NLP models. As TINTA undergoes continuous refinement, expansion, and collaborative efforts across various fields, it is poised to reveal fresh perspectives on the intricacies of Indonesian language variations, paving the way for groundbreaking applications and comprehensive linguistic analysis within the Indonesian context.

## REFERENCES


- Ädel, A. (2020). Corpus Compilation. In *A Practical Handbook of Corpus Linguistics* (pp. 3–24). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_1
- Anthony, L. (2020). Programming for Corpus Linguistics. In *A Practical Handbook of Corpus Linguistics* (pp. 181–207). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_9
- Baayen, R. H., & Linke, M. (2020). Generalized Additive Mixed Models. In *A Practical Handbook of Corpus Linguistics* (pp. 563–591). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_23
- Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education*, *17*(1), 44. https://doi.org/10.1186/s41239-020-00223-0
- Bühlmann, P., Drineas, P., Kane, M., & van der Laan, M. (Eds.). (2016). *Handbook of Big Data*. Chapman and Hall/CRC. https://doi.org/10.1201/b19567

- Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, *52*(4), 460–475. https://doi.org/10.1017/S0261444819000089
- Claridge, C. (2007). Constructing a corpus from the web: message boards. In *Corpus Linguistics and the Web* (pp. 87–108). BRILL. https://doi.org/10.1163/9789401203791_007
- Fletcher, W. H. (2007). Concordancing the web: promise and problems, tools and techniques. In *Corpus Linguistics and the Web* (pp. 25–45). BRILL. https://doi.org/10.1163/9789401203791_004
- Franzosi, R. (2021). What's in a text? Bridging the gap between quality and quantity in the digital era. *Quality and Quantity*, *55*(4), 1513–1540. https://doi.org/10.1007/s11135-020-01067-6
- Gries, S. Th., & Durrant, P. (2020). Analyzing Co-occurrence Data. In *A Practical Handbook of Corpus Linguistics* (pp. 141–159). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_7
- Gries, S. Th., & Paquot, M. (2020). Writing up a Corpus-Linguistic Paper. In *A Practical Handbook of Corpus Linguistics* (pp. 647–659). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_26
- Gut, U. (2020). Spoken Corpora. In *A Practical Handbook of Corpus Linguistics* (pp. 235–256). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_11
- Hirschheim, R. A., & Whitchurch, D. (2023). Big Data, Little Understanding. In *Cambridge Handbook of Qualitative Digital Research* (pp. 43–59). Cambridge University Press. https://doi.org/10.1017/9781009106436.006
- Hoffmann, S. (2007). From web page to mega-corpus: the CNN transcripts. In *Corpus Linguistics and the Web* (pp. 69–85). BRILL. https://doi.org/10.1163/9789401203791_006
- Holmes, D. E. (2017a). *Big Data: A Very Short Introduction* (Vol. 1). Oxford University Press. https://doi.org/10.1093/actrade/9780198779575.001.0001
- Holmes, D. E. (2017b). Big data analytics. In *Big Data: A Very Short Introduction* (pp. 44–58). Oxford University Press. https://doi.org/10.1093/actrade/9780198779575.003.0004
- Holmes, D. E. (2017c). Big data and society. In *Big Data: A Very Short Introduction* (pp. 105–112). Oxford University Press. https://doi.org/10.1093/actrade/9780198779575.003.0008
- Holmes, D. E. (2017d). Big data security and the Snowden case. In *Big Data: A Very Short Introduction* (pp. 90–104). Oxford University Press. https://doi.org/10.1093/actrade/9780198779575.003.0007
- Holmes, D. E. (2017e). Storing big data. In *Big Data: A Very Short Introduction* (pp. 26–43). Oxford University Press. https://doi.org/10.1093/actrade/9780198779575.003.0003
- Holmes, D. E. (2017f). The data explosion. In *Big Data: A Very Short Introduction* (pp. 1–13). Oxford University Press. https://doi.org/10.1093/actrade/9780198779575.003.0001
- Holmes, D. E. (2017g). Why is big data special? In *Big Data: A Very Short Introduction* (pp. 14–25). Oxford University Press. https://doi.org/10.1093/actrade/9780198779575.003.0002
- Hundt, M., Nesselhauf, N., & Biewer, C. (2007). *Corpus Linguistics and the Web*. BRILL. https://doi.org/10.1163/9789401203791
- Jaworska, S. (2016). A comparative corpus-assisted discourse study of the representations of hosts in promotional tourism discourse. *Corpora*, *11*(1), 83–111. https://doi.org/10.3366/cor.2016.0086
- Kehoe, A. (2020). Web Corpora. In *A Practical Handbook of Corpus Linguistics* (pp. 329–351). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_15
- Knight, D., & Adolphs, S. (2020). Multimodal Corpora. In *A Practical Handbook of Corpus Linguistics* (pp. 353–371). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_16
- Lefer, M.-A. (2020). Parallel Corpora. In *A Practical Handbook of Corpus Linguistics* (pp. 257–282). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_12
- Levshina, N. (2020). Conditional Inference Trees and Random Forests. In *A Practical Handbook of Corpus Linguistics* (pp. 611–643). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_25
- Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. In *Corpus Linguistics and the Web* (pp. 7–24). BRILL. https://doi.org/10.1163/9789401203791_003
- McEnery, T. (2000). A new agenda for corpus linguistics - working with all of the world's languages. *Literary and Linguistic Computing*, *15*(4), 403–420. https://doi.org/10.1093/llc/15.4.403
- McEnery, T., & Hardie, A. (2011). *Corpus Linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395
- Mizumoto, A., Plonsky, L., & Egbert, J. (2020). Meta-analyzing Corpus Linguistic Research. In *A Practical Handbook of Corpus Linguistics* (pp. 663–688). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_27
- Newman, J., & Cox, C. (2020). Corpus Annotation. In *A Practical Handbook of Corpus Linguistics* (pp. 25–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_2
- Nugraha, D. S. (2020a). The Comparative Analysis of Syntactic Features Between Indonesian and English Denominal Verbs. *LiNGUA: Jurnal Ilmu Bahasa Dan Sastra*, *15*(1), 65–78. https://doi.org/10.18860/ling.v15i1.7680
- Nugraha, D. S. (2020b). The vlog register in Bahasa Indonesia: an ethnolinguistics study. *International Journal on Language, Research and Education Studies*, *4*(1), 92–103.

- Nugraha, D. S. (2020c). The Information Structure on Newspaper Headlines of the Bahasa Indonesia and the English: a Contrastive Study. *Language and Language Teaching Conference 2019*.
- Nugraha, D. S. (2021a). Makna-makna gramatikal konstruksi verba denominatif dalam bahasa Indonesia. *Bahasa Dan Seni: Jurnal Bahasa, Sastra, Seni, Dan Pengajarannya*, *49*(2), 224. https://doi.org/10.17977/um015v49i22021p224
- Nugraha, D. S. (2021b). Morphosemantic Features of Derivational Affix {Me(N)-} in The Indonesian Denumeral Verb Constructions. *Sirok Bastra*, *9*(2). https://doi.org/10.37671/sb.v9i2.317
- Nugraha, D. S. (2023). The Topical Theme Depicted in the Talks of Education in Indonesian's TEDx : a Systemic Functional Linguistics Approach. *ELT-Lectura*, *10*(1), 21–31. https://doi.org/https://doi.org/10.31849/elt-lectura.v10i1.11905
- O'Keeffe, A., & McCarthy, M. J. (2022). *The Routledge Handbook of Corpus Linguistics*. Routledge. https://doi.org/10.4324/9780367076399
- Paquot, M., & Gries, S. Th. (Eds.). (2020). *A Practical Handbook of Corpus Linguistics*. Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1
- Peter, J.-T., Vilar, D., Deutsch, D., Finkelstein, M., Juraska, J., & Freitag, M. (2023). There's no Data Like Better Data: Using QE Metrics for MT Data Filtering. *Conference on Machine Translation - Proceedings*, 559–575.
- Renouf, A., Kehoe, A., & Banerjee, J. (2007). WebCorp: an integrated system for web text search. In *Corpus Linguistics and the Web* (pp. 47–67). BRILL. https://doi.org/10.1163/9789401203791_005
- Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *6*(2), 159–177. https://doi.org/10.3366/cor.2011.0011
- Rosenbach, A. (2007). Exploring constructions on the web: a case study. In *Corpus Linguistics and the Web* (pp. 167–190). BRILL. https://doi.org/10.1163/9789401203791_011
- Rühlemann, Christoph, & O'Donnell, Matthew Brook. (2012). Introducing a corpus of conversational stories. Construction and annotation of the Narrative Corpus. *Corpus Linguistics and Linguistic Theory*, *8*(2), 313–350. https://doi.org/10.1515/cllt-2012-0015
- Shalabh. (2019). Handbook of Big Data Analytics W. K.HärdleH. H.-S.LuX.Shen eds 2018New YorkSpringer 538 + viii pp., £ 231 ISBN 978-3-319-18283-4. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(4), 1646–1647. https://doi.org/10.1111/rssa.12510
- Th. Gries, S. (2020). Analyzing Dispersion. In *A Practical Handbook of Corpus Linguistics* (pp. 99–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_5
- Xodabande, I., Atai, M. R., Hashemi, M. R., & Thompson, P. (2023). Developing and validating a mid-frequency word list for chemistry: a corpus-based approach using big data. *Asian-Pacific Journal of Second and Foreign Language Education*, *8*(1). https://doi.org/10.1186/s40862-023-00205-5
- Zeldes, A. (2020). Corpus Architecture. In *A Practical Handbook of Corpus Linguistics* (pp. 49–73). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_3