Online version available in : http://arbitrer.fib.unand.ac.id

JURNAL ARBITRER

| 2339-1162 (Print) | 2550-1011 (Online) |

Article

DICO-JALF v.1.0: Diponegoro Corpus of Japanese Learners as a Foreign Language in Indonesia with AI Error Annotation and Human Supervision

Prihantoro[1], Shin'ichiro Ishikawa[2], Tanjun Liu[3], Zaki Ainul Fadli[4], Elizabeth Ika Hesti Aprilia Nindia Rini[5], Catur Kepirianto[6]*

[1,4,5,6]Faculty of Humanities, Universitas Diponegoro, Semarang, Indonesia
[2]IPHE/Graduate School of Intercultural Studies, Kobe University, Kobe, Japan
[3]Department of Applied Linguistics, Xi'an Jiaotong-Liverpool University, Suzhou, China

Submission Track

Received: May 11, 2025
Final Revision: August 1, 2025
Accepted: August 14, 2025
Available Online: September 25, 2025

Keywords

Corpus linguistics, error annotation, AI, ChatGPT, Indonesia, JFL

Correspondence

*E-mail: prihantoro@live.undip.ac.id

ABSTRACT

There is a growing body of research in using AI for corrective feedback in foreign language teaching. However, few studies have specifically addressed the accuracy of AI analysis in learner corpus research. This study aims to create an AI-annotated corpus whose data were obtained from learners of Japanese as a Foreign Language (JFL) in Indonesia with human supervision; branded it as DICO-JALF v.1.0. The aim is to measure to what extent ChatGPT accurately annotates errors. A task was first administered to collect corpus data and metadata to build the corpus. The corpus was error-annotated using ChatGPT 4.0. Human annotators manually supervised the accuracy of AI-generated annotations. Regarding errors committed by learners, it is observed that incorrect lexical choices and forms dominate the cause of errors, while underuse and overuse are minimal. It can be concluded that ChatGPT demonstrated an average accuracy of 70% correct identification of errors. Regarding error rate, the verb is the category where errors are most frequent, which maybe driven by its conjugation, a feature absent in Indonesian, the L1 of the students. This suggests that Indonesian learners' acquisition of Japanese verbs needs greater emphasis. As compared to other similar studies, this is relatively low. However, it can be argued that one factor determining the accuracy of ChatGPT annotations, or any other LLM-based tool, is the complexity of the annotation scheme they adhere to. The corpus have been made available for download. The annotations shall be readable by a corpus query system that reads XML tags. This corpus serves as a foundational resource for future research on AI-assisted error analysis in JFL learning contexts in Indonesia.

I. INTRODUCTION

Learner corpus is one of the resources to study how to tackle language learners' challenges. Learner corpora have the potential to describe learners' linguistic development and errors, as attested in Perez-Paredez and Mark (2022), and Forti (2023), among others. However, the creation of error-annotated learner corpora is usually performed manually because an automatic error tagger is absent. LLMs (Large Language Models), which power various AI (Artificial Intelligence) tools, provide an alternative to automating the annotation process, which allows for less human intervention and a faster process. However, the accuracy and reliability of AI tools in the area still need to be assessed. This project describes a learner corpus creation, whose error annotations were conducted automatically by an AI tool,

DOI: https://doi.org/10.25077/ar.12.3.274-288.2025

with an assessment of the AI's performance. This background translates to the following objectives and potential contributions.

The first objective 1) was to create DICO-JALF v.1.0, a grammatically error-annotated corpus from Indonesian students learning Japanese as a foreign language at various proficiency levels. This corpus will offer a comprehensive data set of learners' performance. *The second objective* 2) was to describe how accurately ChatGPT 4.0 (OpenAI, 2024) annotated errors in essays written in Japanese. This will allow ChatGPT's performance to be assessed when analysing non-English data.

In Indonesia, in addition to English, Japanese is another popular foreign language to learn[1]. Among other purposes, Japan attracts many Indonesian migrant workers and students to work, pursue higher education, or a combination thereof, as shown in Umoro (2023) and Budianto (2023). Note that Japanese as a foreign language (JFL) is popular not only in Indonesia, as shown in Table 1.

Table 1. Number of Learners of JFL Across the World[2]

No	Countries	Number of learners
1	China	1057318
2	Indonesia	711732
3	Korea	470334
4	Australia	415348
5	Thailand	183957

Numerous studies have been conducted in the context of Indonesian students learning Japanese, such as Safama and Diner (2022), Barus and Pujiono (2021), among many others. These studies focus on analysing errors committed by learners of Japanese as a foreign language in Indonesia, but unlike this project, they focus on certain linguistic unit such as particle or sentence structure. However, some handicaps were observed. First, none of these studies were corpus-based. Second, if their data is considered Japanese learner corpora, none of these corpora are publicly accessible. Such a corpus would be helpful for replication studies, cross-validation, and comparison purposes, among many others. Third, the analytic schemes used in the aforementioned studies targeted specific errors, such as passive voice and particles, instead of overall errors at different level of linguistic structures. Using their analytic scheme for the annotation can leave loopholes, as it does not match the purpose of fully annotating errors in a Japanese learner corpus. Hasibuan and Arfianty (2019) conducted a similar study, but their tagset is incomprehensive. For instance, it includes an analysis of 'incorrect word use' without giving further details.

So far, three comprehensive error type tagsets (Koyama et al., 2023; Pavlovič, 2020; Yang and Akahori, 2013) have been identified, which may suit the purpose. Compared to Pavlovic (2020), Koyama et al.'s (2023) tagset is better regarding tagset documentation. For example, Pavlovic's tagset provides references, but no labels or examples. Conversely, labels, references/descriptions, and examples in Koyama's tagset can be observed. In addition, Pavlovic's tagset adopts more technical linguistic terms. For this current study, it is aimed to use a tagset that would allow us to reach a more general audience; thus, Koyama's full tagset, which can be observed in Table 2, was preferred. Yang and Akahori's (2013) tagset was not considered. This tagset is a subset of tags presenting the aforementioned tags but is not as comprehensive as the other two.

While Koyama's tagset is comprehensive, there is no tag to identify the absence of a required item. This is a common mistake made by learners, which should be tagged. For example, in (1), the verb *shimashita* をしました 'do (past and polite)' is absent, thus causing an error to occur. The support verb should be immediately after *sakka* サッカー 'soccer', as it must surface at the sentence-final position to allow the sentence to be grammatically correct.

(1) こどものときまいにちともだちといっしょにサッカー()[3]。
Kodomo-no-toki-mainichi-tomodachi-to-issho-ni-sakkaa()
kid-GEN-time-everyday-friend-COM-together-LOC-soccer()
'When I was a child, every day, together with my friends, we (played) soccer'

A slight adjustment to Koyama et al.'s tagset was commited by incorporating Pavlovic's error cause tagset as the second layer tagset, whose categories are shown in Table 3.

Pavlovic's error type tagset consists of labels, without descriptions or examples. To complement this, some examples and labels were added. This

[1] https://www.qiteplanguage.org/detail/news/SEA-QIL-dan-UPI-teliti-preferensi-bahasa-siswa-di-Asia-Tenggara

[2] https://www.jpf.go.jp/e/project/japanese/survey/result/dl/survey2021/All_contents_r2.pdf

[3] The empty brackets with zero element () refers to missing item.

Table 2. Error type tagset (modified from Koyama (2023)

Tag	Description	Example {incorrect form → correct form}
ADJ	Adjective selection error	そんなことは一番｛大切→重大な｝欠点だと思うでしょう。 Sonna koto wa ichiban {taisetsu → juudaina} ketten da to omou deshou. You might think that it is a {serious} weakness
ADV	Adverb selection error	そして、自分の開発とテストの仕事を｛よく→しっかり｝完遂しました。 Soshite, jibun no kaihatsu to tesuto no shigoto o {yoku → shikkari} kansui shimashita. and i have {thoughtfully}completed my development and testing works
AUX	Auxilliary selection error	私は日本語が大好き｛ます→です｝。 Watashi wa nihongo ga daisuki {masu → desu}. I {like} Japanese very much
CONJ	Conjunction selection error	｛しかし→そして｝、私の一番の夢はタイへ行って、タイ語で現地の人と交流することです。 {Shikashi → Soshite}, watashi no ichiban no yume wa Thai e itte, Thai go o genchi no hito to kouryuu suru koto desu. {and} my biggest dream is to go to Thailand and interact with local people in Thai.
DET	Determiner selection error	｛あの→その｝化学会社の名前はチッソ株式会社だった。 {Ano → sono} kagaku no kaisha no namae wa Chisso kabushiki gaisha datta. {that} chemical substance factory is called Chico incorporated company
NOUN	Noun selection error	その人はソフトウェア｛権力→権利｝を持っている。 Sono hito wa sofuto wea {kenryoku→ kenri } o motte iru. The person has the {posession right} of the software
PART	Particle selection error	今日、雷の音｛に→で｝起きました。 Kyou, kaminari no oto {ni → de} okimashita. I woke up {because} of thunder
PRON x	Pronoun selection error	まあ、今回の話は｛そこ→ここ｝までにします。 Maa, konkai no hanashi wa {soko → koko} madeni shimasu. Alright, we {here} cease our conversation
PUNCT	Punctuation selection error	一緒に歌を歌いました｛、→。｝ご飯を食べて、買い物をしました。 Isshoni uta o utaimashita {, → . } gohan o tabete, kaimono o shimashita. We sang together {. → ,} then went shopping
VERB	Verb selection error	朝ご飯を｛飲み→食べ｝ました。 i {ate} breakfast Asa gohan o {nomi → tabe} mashita. I {eat} breakfast
ADJ : INFL	Adjective conjugation error	突然とても｛寂しい→寂しく｝感じ始めました。 Totsuzen totemo {sabishii → sabishiku} kanjihajimemashita. Suddenly, i felt very {lonely}
AUX : INFL	Auxilliary conjugation error	しゅくだいはとてもつまら｛ない→なく｝てとてもむずかしいです。 Shukudai wa totemo tsumara {nai → naku } te totemo muzukashii desu. The homework is extremely {boring} and difficult
VERB : INFL	Verb conjugation error	しかし、ネットで探すと、ぜんぜん｛見つかれ→見つから｝ないでしょ！ Shikashi, netto de sagasu to, zensen {mitsukare → mitsukara} nai desho ! {but, if you search on the internet, it is completely {not found} right?
SPELL	Spelling error	国内の｛メデイア→メディア｝も管理されて過激な言論はいっさい禁止されています。 Kokunai no {medeia→ media} mo kanri sarete kagekina genron wa issai kinshi sarete imasu. Domestic {media} has been regulated, and all extreme speech is completely forbidden
VERB : TENSE	Tense usage error	知っ｛た→ている｝人気ウェブがあるから、ぜひお知らせください。 Shit {ta→ te iru} ninki webbu ga aru kara, zehi oshirase kudasai. If you {know} any popular website, please inform me
WO	Word order error	日本語｛四級検定→検定四級｝合格！ Nihongo {shikyuu kentei → kentei shikyuu} goukaku ! I just passed {level 4 of Japanese Proficiency Test}
OTHER	Other error	数秒後、ベランダーは｛見えるような速度を立て→見る見るうちに立って｝いきます。 Suubyougo, berandaa wa {mieru youna sakudo o tate → miru miru uchi ni tatte} ikimasu. In the next few seconds, the {veranda} swiftly stood before me

Table 3 Error cause tagset (modified from Pavlovic (2020) CC-BY)

No	Tag	Label	Example {incorrect form → correct form}
1	Wrong choice	WRONG_CHOICE	ゲーム｛を→が｝好きです。
2	Lack of use	LACK_USE	ゲーム｛Ø→が｝好きです。
3	Form error	FORM_ERROR	この考えはもう忘れて｛しました→しまいました｝。
4	Overuse	OVERUSE	あの時からよく「HyannaNatsu」のスピードペイント動画を見て、Hyanna先生の描き方、線画の描きとか、ペイントする｛の→Ø｝方法などを勉強してました。

is one modification. Second, 'redundancy' is replaced with an 'overuse' category. It is assumed that overuse can include redundancy, but not vice versa, because the latter is restricted to repeating one item (with another identical one), which causes the sentence to be ill-formed. To clarify label 1 and 3 'Wrong choice' means that the form is correct in isolation, but contextually erroneous. Meanwhile, 'FORM_ERROR' means that the learner attempted to use a correct item, but used the wrong form (e.g., wrong inflection, conjugation, incomplete word, among others). Overall, this tagset complements the previous tagset, which mixes form and POS segment errors. The tagset focus on explaining the cause of errors, neutral from POS segment where the errors are commited.

Compare (2) and (3). In sentence (2), it is observed that the repetition of *toki* follows *sono*. The latter is unnecessary and can therefore be categorised as either redundancy or overuse. While the two concepts are interchangeable in (2), in (3), only 'overuse' fits. This is because the erroneous segment の no (genitive marker) is not repeated from any segment. Its use is unnecessary because, without it, the sentence is already acceptable.

(2)高校生でコロナウイルスがあった時その時人々はみんな家にいました。１６５

Koukousei-de-korona-virusu-ga-atta-toki-sono-toki-hito-bito-wa-minna-ie-ni-imashita

Student.high-school-CONJ-COVID19-SUBJ.CASE-EXT-time-that-people-TOP-all-house-LOC-EXT

'When I was in high school and there was a coronavirus pandemic, everyone stayed at home'

(3)()LITTLEMIXのが一番好きです。

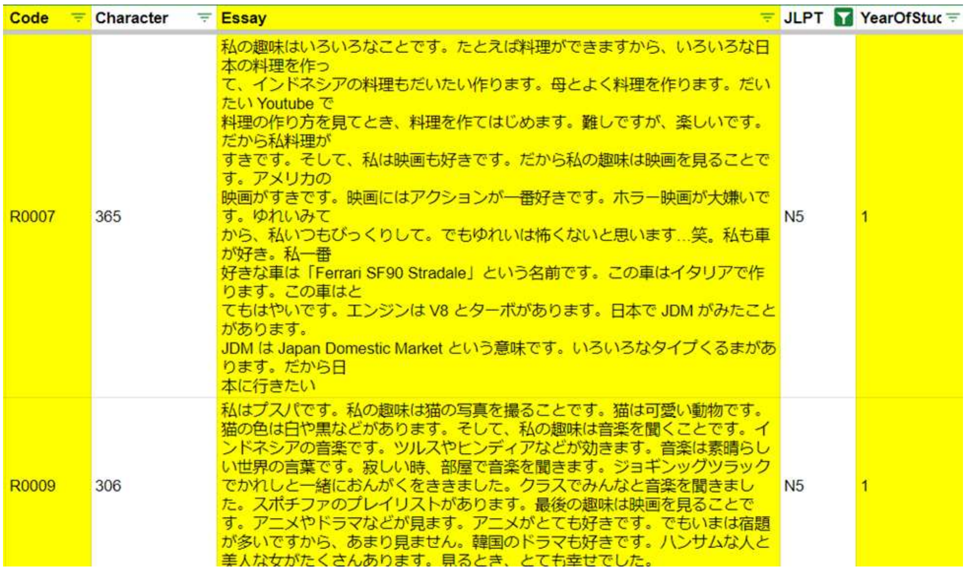
()LTTLEMIX-no-ga-ichiban-suki-desu.
()LITTLEMIX-GEN-PS-most-like-COP.
'(I) really like little mix'

For the first aim, the literature review can be concluded as follows. First, the need to construct the corpus targeted in this project is justified by the absence of a Japanese essay corpus written by learners of JFL whose L1 is Indonesian. Second, a review shows that the error analysis schemes proposed by other scholars are insufficient, as the study aims to annotate multiple errors instead of certain ones. Koyama et al.'s error types tagset and Pavlovic's error cause tagset were then combined into a two-layer error tagset,.

Artificial Intelligence (AI) is a generic cover term for systems that can implement human intelligence tasks. A recent trend in AI is the use of Large Language Models (LLMs), language models trained on a large amount of data to interact using human instead of machine language (e.g. programming languages). In addition to producing texts (creative writing, translation, proofread texts), LLMs can also generate images, voices, and video, among others. While studies on the use of AI in foreign language education are numerous (e.g. Karataş et al., 2024; Pan et al., 2024), its application in corpus linguistics, particularly in corpus building, is understudied. Among the few that exist, I first highlighted Yu et al. (2024) who assessed the potential of LLM for corpus pragmatic annotation, comparing ChatGPT 3.5 and 4.0. They argue that ChatGPT 4.0 is better, and its performance (Precision-Recall and F1 Score) is impressively above 90% for all speech acts. Based on this finding, ChatGPT 4 is used, whose performance is better.

However, Yu et al.'s (2024) study is different from this project in at least two respects. First, Yu et al.'s (2024) study targeted English instead of Japanese. The size and availability of LLM's training data sets for English and Japanese are not comparable. Second, this project does not aim to apply AI to conduct pragmatic annotations, as Yu et al. did.

Compared to Yu et al.'s study, Poole and Coss' (2024) study is more similar to ours. Pool and Coss (2024) studied how ChatGPT to apply a writing rubric to essay correction tasks and evaluated is performance. While Poole and Coss (2024) argue that ChatGPT can serve as a valuable tool for L2



Code	Character	Essay	JLPT	YearOfStuc
R0007	365	私の趣味はいろいろなことです。たとえば料理ができますから、いろいろな日本の料理を作って、インドネシアの料理もだいたい作ります。母とよく料理を作ります。だいたい Youtube で料理の作り方を見てとき、料理を作てはじめます。難しですが、楽しいです。だから私料理がすきです。そして、私は映画も好きです。だから私の趣味は映画を見ることです。アメリカの映画がすきです。映画にはアクションが一番好きです。ホラー映画が大嫌いです。ゆれいみてから、私いつもびっくりして。でもゆれいは怖くないと思います...笑。私も車が好き。私一番好きな車は「Ferrari SF90 Stradale」という名前です。この車はイタリアで作ります。この車はとてもはやいです。エンジンは V8 とターボがあります。日本で JDM がみたことがあります。JDM は Japan Domestic Market という意味です。いろいろなタイプくるまがあります。だから日本に行きたい	N5	1
R0009	306	私はプスパです。私の趣味は猫の写真を撮ることです。猫は可愛い動物です。猫の色は白や黒などがあります。そして、私の趣味は音楽を聞くことです。インドネシアの音楽です。ツルスやヒンディアなどが効きます。音楽は素晴らしい世界の言葉です。寂しい時、部屋で音楽を聞きます。ジョギンッグツラックでかれしと一緒におんがくをききました。クラスでみんなと音楽を聞きました。スポチファのプレイリストがあります。最後の趣味は映画を見ることです。アニメやドラマなどが見ます。アニメがとても好きです。でもいまは宿題が多いですから、あまり見ません。韓国のドラマも好きです。ハンサムな人と美人な女がたくさんあります。見るとき、とても幸せでした。	N5	1

Fig 1. Sample results: respondent code, essay, JLPT score, year of study

writing assessment, it fails to reach a desirable threshold. The objective of the study and this study is similar: applying some kind of rubric to L2 writing assessment. The difference is that, this project is more specific in that an annotated error analysis scheme, not writing in general is applied. Another difference is the target language, which is not English but Japanese. In a different language, the analytic scheme may differ due to the language typology. For instance, 'merge' and 'split' are categories[4] of errors present in Arabic, as shown in Alrehli and Alhotahli's (2025) AI-Assisted error analysis, but not in English.

The aforementioned critical reviews translate to the novelty of this project. This is the first learner corpus whose data is purely obtained from Indonesian students learning Japanese. The architecture of the corpus and the data collection procedure are transparent. The corpus is free to access. This allows for replicability, reproducibility, and data expansion under the same protocol. Error annotations and existing metadata can lead to corpus-based learner profiling. In terms of studies using AI, many studies have focused on using AI tools, while very few studies have specifically addressed the accuracy of AI tools in conducting error annotation. The results of this study may be useful for improving automated language assessment tools.

II. METHODS

Corpus data were collected from respondents in the Department of Japanese Language Universitas Diponegoro (UNDIP) for several reasons. First, UNDIP is one of the few universities whose Japanese language departments are nationally accredited as A-grade. While the data only cover students from one university, it is argued that this is a sensible starting point before improving representativeness in subsequent studies. Starting with relatively small data is one of the approaches to solidifying the data collection and analysis protocol before expanding to a larger data set.

Approximately 300 students in years 1–3 (year 2021-2023), were asked to write a short essay ranging from 300–600 characters (Katakana, Hiragana, Kanji, or a combination thereof) on a single topic, 'my hobby' a single topic, 'my hobby' (from June-July 2024). While obvously expandable, and this is considered to be sufficient as a starting point. It shall be addressed in future studies by eliciting and collecting essays on various topics. The consideration for a range of short essays is that the majority of students in year one started as absolute beginners (e.g. minimum proficiency in writing any Japanese characters at all), even though by the time data were collected, they had already learned Japanese for almost one semester.

The students were asked to write an essay in 75 minutes and were not allowed to use any AI writing assistance software or dictionaries. Upon completing their essays, they were requested to complete a questionnaire. This is aimed to obtain their metalinguistic information (year of study, JLPT score, sex, among others). The results were saved in spreadsheet files, as shown in Figure 1.

[4] Note that the remaining categories, however, are quite generic such as semantics, morphology, syntax or punctuation

The essays and questionnaires were carefully checked to ensure that all essays complied with the character limit, all required metadata information was supplied, and the respondents clearly expressed giving consent for data access and publication. Data that did not meet these requirements (3 essays) were excluded from the sample. The returned essays were randomly sampled, and stratified by year of study: 30% from each year (see Table 4). The sample comprised 107 essays (around 33% of the population), which would form the corpus.

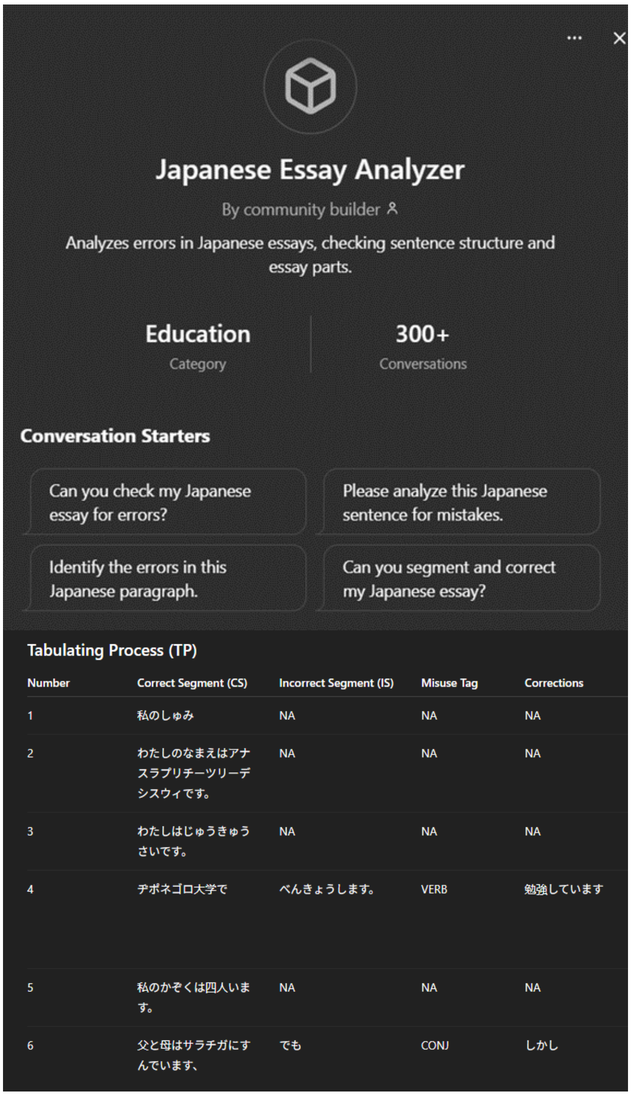
Table 4. Quick summary of the sample[5]

No	Information	Year 1	Year 2	Year 3
1	Essay	35	36	36
2	Average characters	372	410	522
3	JLPT/No-JLPT	5/30	15/20	30/5
4	Male/Female	15/20	13/22	9/26
5	Studied Japanese before college (Y/N)	6/29	2/33	20/10

AI analysis and supervised corpus data creation

An assessment was implemented on these essays with help from ChatGPT 4.0, which was reported to perform better than its predecessor, ChatGPT 3.5, as shown by Holland (2023) and Massey et al. (2023). The prompts requested ChatGPT to identify each incorrect segment and assign error types and cause tags. The prompts were installed on ChatGPT's interface (see Japanese Essay Analyser in Fig. 2) and applied them to each essay.

Unlike Park (2019) who compared ChatGPT and human works for error analysis, in this study, the results of ChatGPT analysis were reviewed by human reviewers. The human reviewers[6] reviewed ChatGPT responses using a 4-parameter metric, namely, Correct Segment[7] (CS), Incorrect Segment (IS), Error Type (ET), and Error Cause (EC), as shown in Table 5. The human reviewers validated whether or not the values given for these four parameters were accurate, using Boolean values (T=True, F=False), as shown in Table 5.



Japanese Essay Analyzer

By community builder

Analyzes errors in Japanese essays, checking sentence structure and essay parts.

Education
Category

300+
Conversations

Conversation Starters

Can you check my Japanese essay for errors?

Please analyze this Japanese sentence for mistakes.

Identify the errors in this Japanese paragraph.

Can you segment and correct my Japanese essay?

Tabulating Process (TP)

Number	Correct Segment (CS)	Incorrect Segment (IS)	Misuse Tag	Corrections
1	私のしゅみ	NA	NA	NA
2	わたしのなまえはアナスラプリチーツリーデシスウィです。	NA	NA	NA
3	わたしはじゅうきゅうさいです。	NA	NA	NA
4	ヂポネゴロ大学で	べんきょうします。	VERB	勉強しています
5	私のかぞくは四人います。	NA	NA	NA
6	父と母はサラチガにすんでいます、	でも	CONJ	しかし

Fig 2. ChatGPT's Installed prompt [Japanese Essay Analyzer] and its output sample

Once the evaluation was concluded, each essay was converted , and its corresponding metadata information, into an XML document to be readable in the user's preferred corpus query system. The raw corpus in case users do not need error annotation attributes is also provided. The corpus is named DICO-JALF[8] v 1.0. This procedure helped to fulfil the first objective of this study.

For this paper's visualisation and analysis purposes in this paper, Sketch Engine was used to index the XML-annotated corpus. The preference for Sketch Engine is because it includes a Japanese POS tagger by default.

The information from the corpus was used to measure the accuracy of ChatGPT's annotation for each parameter in the metric and also overall accuracy. Accuracy is operationally defined in

[5] The slash in each year column correspond to the information. for instance 5/30 in line 4 year 1 means that of 35 students sampled in year 1, 5 have JLPT certificates while the other 30 do not have any JLPT certificate.

[6] Japanese lecturers/tutors, 2.5 years or more teaching experience, minimum proficiency JPLT N2

[7] Questions asked to reviewers => CS: Is there any error in this segment? IS:Is there any error in this segment? ET: Is the assigned error tag correct? EC: is the error correction accurate? If yes, write T, if not write F.

[8] DICO-JALF = Diponegoro Corpus of Japanese Learners as a Foreign Language

Table 5. Sample evaluation metric

No	CS	*EV CS*	IS	*EV IS*	*ET*	*EV ET*	EC	*EV EC*
1	この趣味は子供の時から今まで	T	続きます。	*T*	VERB_INFL	*T*	Form error	*T*
2	それから私のしゅみはホラーオンラインマンガを	T	よみます。	*T*	VERB	*F*	Form Error	*T*
3	たくさんオンラインゲームを	T	あそびます。	*T*	VERB	*T*	Form Error	*F*

Table 6. Quick summary of the corpus

No	CS	FULL	EVAL
1	Token	70070	19223
2	Text (Essay)	325	107
3	Metadata attribute	Yes	Yes
4	POS-tagged (MeCab)	Yes	Yes
5	GPT 4.0-Error-annotated	No	Yes
6	Human evaluation	No	Yes

this project as the proportion of TRUE values compared to all values (TRUE+FALSE). The reference was made to the average value of all four parameters for overall accuracy (cf. Table 5, and Tabel 7). In addition, the likelihood of each error type label assigned by ChatGPT being correct was also measured. This technically translates to the proportion of each error type label assigned by ChatGPT evaluated as 'TRUE' by human annotators. The accuracy of each error type was then measured. The annotators assigned the correct error type label for each mistake made by students following Koyama's modified tagset. The proportion of matching error-type labels given by ChatGPT was calculated. This helped to fulfil the second objective of this study. Unlike Prihantoro (2022), this project does not use precision-recall as an evaluation measure because the tags are unambiguous, while precision-recall is an evaluative measure typically used in information retrieval systems, beyond this study's coverage. For unambiguous tags, accuracy is a better fit, as commonly used on other projects such as Prihantoro (2025), Pandey (2002), or Thewissen (2013).

III. RESULTS

Objective 1

Architecture

The first research objective, to create DICO-JALF v.1.0, has been fulfilled. Here, two XML versions of DICO-JALF were created: RAW and ANNOTATED. The RAW version includes all essays (325), while the ANNOTATED version includes a sample (107[9]), as shown in Table 6. The latter was enhanced with ChatGPT error annotation and human evaluation. These versions are freely available for download[10] and use in users' preferred corpus query systems.

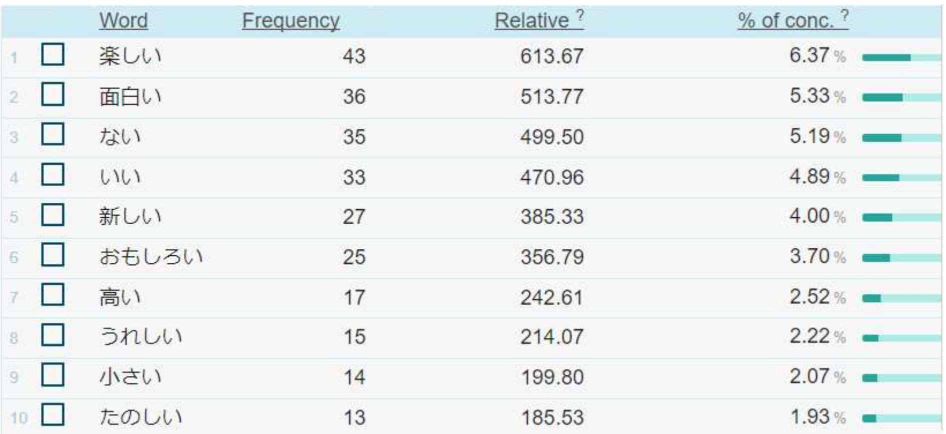
Regardless, all essays with their corresponding metadata were indexed in SE. As POS tags and metadata are present, users can perform POS-based searches with metadata restrictions, as shown in Figure 3. It shows adjectives used by female students ranked by frequency.

The corpus created corpus fulfils the first aim of this research, which is to create an error-annotated corpus from Indonesian students learning Japanese as a foreign language at various levels. Unlike Imamura et al.'s (2012) experiment in which pseudo data were used, in this study the data used were obtained authentically from human learners of Japanese as a foreign language whose native or dominant language is Indonesian.

If specifically compared to other learner corpora in Japanese, such as the International Corpus of Japanese as a Second Language (IJAS), Natsume (Nishina et al., 2014), the learner corpus created for Imamura et al.'s (Imamura et al., 2014) experiment, or NAIST lang 8 (Kasahara et al., 2011), this corpus is smaller. While this may be considered a handicap, it is argued to be a sensible starting point.

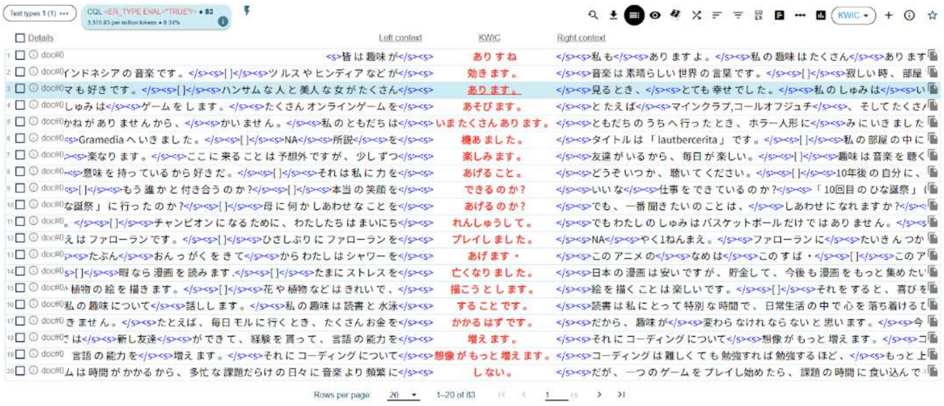
[9] This is around 30% of all essays, a manageable amount considering the time and financial constraints. The commitment is to expand this once more finance and time resources are available.

[10] https://drive.google.com/drive/folders/11QREjUrMjARfKOei5m-lPpVubjKhQMIar?usp=drive_link



		Word	Frequency	Relative ?	% of conc. ?
1	□	楽しい	43	613.67	6.37 %
2	□	面白い	36	513.77	5.33 %
3	□	ない	35	499.50	5.19 %
4	□	いい	33	470.96	4.89 %
5	□	新しい	27	385.33	4.00 %
6	□	おもしろい	25	356.79	3.70 %
7	□	高い	17	242.61	2.52 %
8	□	うれしい	15	214.07	2.22 %
9	□	小さい	14	199.80	2.07 %
10	□	たのしい	13	185.53	1.93 %

Fig 3. Frequency of adjectives used by female students in DICO-JALF's SE (Generated by Word List tool)



Text types 1 (1) •••

CQL <ER_TYPE EVAL="TRUE"/> • 83
3,370.83 per million tokens • 0.34%

	□ Details	Left context	KWIC	Right context
1	□ ⓘ doc#0	<s>皆 は 趣味 が</s><s>	あり す ね	</s><s>私 も</s><s>あり ます よ 。</s><s>私 の 趣味 は たくさん</s><s>あり ます
2	□ ⓘ doc#0	インドネシア の 音楽 です 。</s><s>[ ]</s><s>ツ ルス や ヒンディア など が</s><s>	効き ます 。	</s><s>音楽 は 素晴らしい 世界 の 言葉 です 。</s><s>[ ]</s><s>寂しい 時 、 部屋
3	□ ⓘ doc#0	マ も 好き です 。</s><s>[ ]</s><s>ハンサム な 人 と 美人 な 女 が たくさん</s><s>	あり ます 。	</s><s>見る とき 、</s><s>とても 幸せ でした 。</s><s>私 の しゅみ は</s><s>い
4	□ ⓘ doc#0	しゅみ は</s><s>ゲーム を し ます 。</s><s>たくさん オンラインゲーム を</s><s>	あそび ます 。	</s><s>と たえ ば</s><s>マインクラブ,コールオフジュチ</s><s>、 そして たくさ
5	□ ⓘ doc#0	かね が あり ませ ん から 、</s><s>かい ませ ん 。</s><s>私 の ともだち は</s><s>	いま たくさん あり ます 。	</s><s>ともだち の うち へ 行っ た とき 、 ホラー人形 に</s><s>み に いき まし た
6	□ ⓘ doc#0	<s>Gramedia へ いき まし た 。</s><s>[ ]</s><s>NA</s><s>所説</s><s>を</s><s>	機あ まし た 。	</s><s>タイトル は 「lautbercerita」 です 。</s><s>[ ]</s><s>私 の 部屋 の 中 に
7	□ ⓘ doc#0	><s>楽なり ます 。</s><s>ここ に 来る こと は 予想外 です が 、 少し ずつ</s><s>	楽しみ ます 。	</s><s>友達 が いる から 、 毎日 が 楽しい 。</s><s>[ ]</s><s>趣味 は 音楽 を 聴く
8	□ ⓘ doc#0	<s>意味 を 持っ ている から 好き だ 。</s><s>[ ]</s><s>それ は 私 に 力 を</s><s>	あげる こと 。	</s><s>どうぞ いつ か 、 聴い て ください 。</s><s>[ ]</s><s>10年後 の 自分 に 、
9	□ ⓘ doc#0	<s>[ ]</s><s>もう 誰 か と 付き合う の か ?</s><s>[ ]</s><s>本当 の 笑顔 を</s><s>	できる の か ?	</s><s>いい な</s><s>仕事 を でき ている の か ?</s><s>「10回目 の ひな誕祭」 (
10	□ ⓘ doc#0	な誕祭」 に 行っ た の か ?</s><s>[ ]</s><s>母 に 何 か しあわせ な こと を</s><s>	あげる の か ?	</s><s>でも 、 一番 聞き たい の こと は 、</s><s>しあわせ に なれ ます か ?</s><
11	□ ⓘ doc#0	。</s><s>[ ]</s><s>チャンピオン に なる ために 、 わたしたち は まいにち</s><s>	れんしゅうし て 。	</s><s>でも わたし の しゅみ は バスケットボール だけ で は あり ませ ん 。</s><s
12	□ ⓘ doc#0	え は ファローラン です 。</s><s>[ ]</s><s>ひさしぶり に ファローラン を</s><s>	プレイし まし た 。	</s><s>NA</s><s>やく1ねんまえ 。</s><s>ファローラン に</s><s>たいき ん つか
13	□ ⓘ doc#0	><s>たぶん</s><s>おん っ がく を き て</s><s>から わたし は シャワー を</s><s>	あげ ます ・	</s><s>この アニメ の</s><s>なめ は</s><s>この す ば ・</s><s>[ ]</s><s>この ア
14	□ ⓘ doc#0	s>[ ]</s><s>暇 なら 漫画 を 読み ます .</s><s>[ ]</s><s>たまに ストレス を</s><s>	亡くなり まし た 。	</s><s>日本 の 漫画 は 安い です が 、 貯金し て 、 今後 も 漫画 を もっと 集め たい
15	□ ⓘ doc#0	植物 の 絵 を 描き ます 。</s><s>[ ]</s><s>花 や 植物 など は きれい で 、</s><s>	描こう と し ます 。	</s><s>絵 を 描く こと は 楽しい です 。</s><s>[ ]</s><s>それ を する と 、 喜び を
16	□ ⓘ doc#0	私 の 趣味 について</s><s>話しし ます 。</s><s>私 の 趣味 は 読書 と 水泳</s><s>	する こと です 。	</s><s>読書 は 私 に とっ て 特別 な 時間 で 、 日常生活 の 中 で 心 を 落ち着ける て
17	□ ⓘ doc#0	き ませ ん 。</s><s>たとえば 、 毎日 モル に 行く とき 、 たくさん お金 を</s><s>	かかる はず です 。	</s><s>だから 、 趣味 が</s><s>変わら なけれ なら ない と 思い ます 。</s><s>今
18	□ ⓘ doc#0	は</s><s>新し友達</s><s>が でき て 、 経験 を 貰っ て 、 言語 の 能力 を</s><s>	増え ます 。	</s><s>それ に コーディング について</s><s>想像 が もっと 増え ます 。</s><s>コ
19	□ ⓘ doc#0	言語 の 能力 を</s><s>増え ます 。</s><s>それ に コーディング について</s><s>	想像 が もっと 増え ます 。	</s><s>コーディング は 難しく て も 勉強すれ ば 勉強する ほど 、</s><s>もっと 上
20	□ ⓘ doc#0	ム は 時間 が かかる から 、 多忙 な 課題だらけ の 日々 に 音楽 より 頻繁 に</s><s>	し ない 。	</s><s>だが 、 一つ の ゲーム を プレイし 始め たら 、 課題 の 時間 に 食い込ん で

Rows per page 20 ▾ 1–20 of 83 1

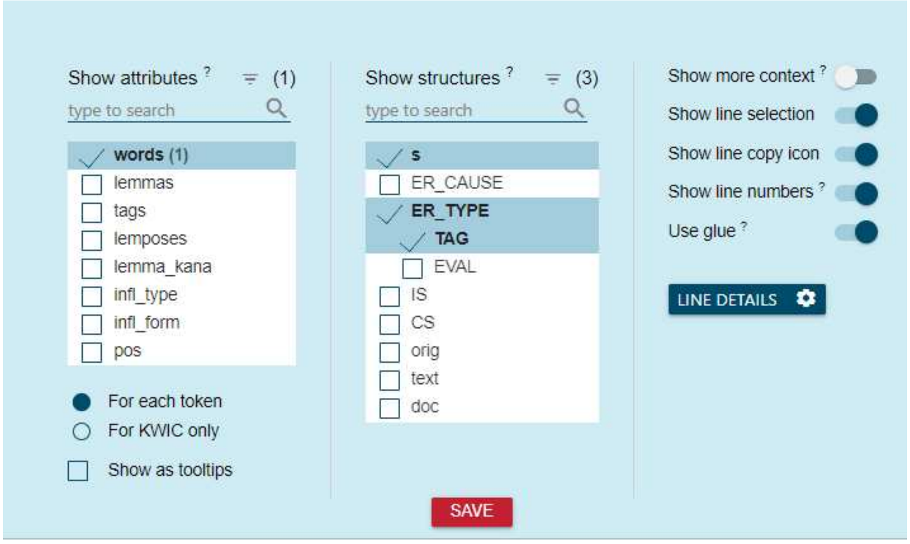
Fig 4. Concordance lines from <ER_TYPE EVAL="TRUE"/> with VERB restriction in ER_TYPE Text Types in DICO-JALF's SE (Generated by Concordance tool)

In terms of the contribution in the context of Japanese learners whose L1 is Indonesian, the data were specifically obtained from native speakers of Indonesian, an aspect missing from other learner corpora of Japanese. There is indeed a small subset of IJAS data that came from Indonesian students. However, the number of students for the experiment was just 50. In contrasts, for this project, data from 325 students (107 error annotated) was collected , far outweighing IJAS respondents. Also IJAS collect data of students from various countries. Conversely, the focus of this study is on Indonesian students. Thus, this corpus can characterise errors generated by Japanese learners as a foreign language in Indonesia better. In future years, the plan is to expand the corpus size by covering all the universities in Indonesia in which Japanese study programmes are offered.

Corpus search based on error tags

Users can search the corpus based on error type or cause tags using the following CQL format <ER_TYPE TAG="VERB"/> (to search for verb category errors) or <ER_CAUSE TAG="WC"/> (to search for wrong choice errors). Note that, in aim 2, it is argued that ChatGPT is not entirely accurate. Thus, users can also incorporate human evaluation information. An example is <ER_TYPE EVAL="VERB"> with restriction to TRUE in the ER_TYPE Text Types. This means that it restricts the search for verb errors that have been evaluated as TRUE by human evaluators. This increases the likelihood that users will get the desired outcome (see Fig. 4). The figure below shows concordance lines whose nodes are error-tagged with 'VERB' and evaluated TRUE by human annotators, presented alongside their left and right contexts.

In some situations, users might want to perform a more underspecified search, such as showing all erroneous segments, error types, and causes (and, in some cases, human evaluation). While, by default, the segment that fits the query is visible in concordances, tags will only be visible based on users' actions. Their visibility might be helpful to users wishing to identify types of errors and causes of errors (as well as their evaluation). Figure 6 shows that users can checkmark the attributes they want to show, while Figure 5 shows how these elements are visually presented in



Show attributes ? (1)
type to search
☑ words (1)
□ lemmas
□ tags
□ lemposes
□ lemma_kana
□ infl_type
□ infl_form
□ pos

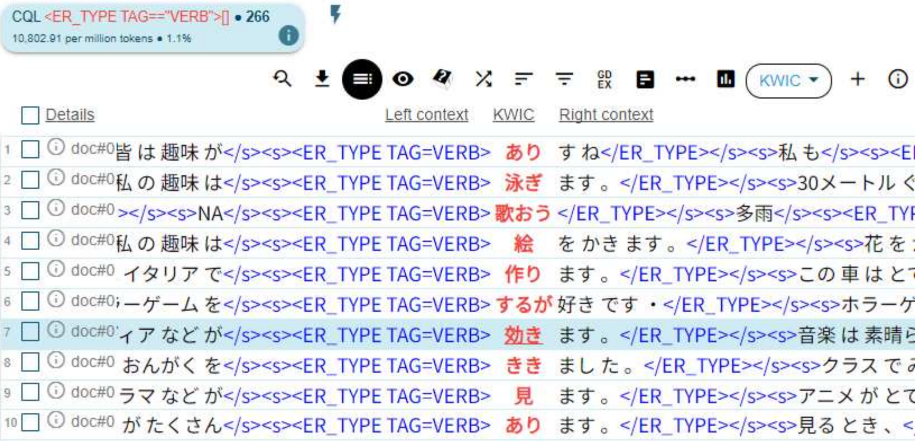
● For each token
○ For KWIC only

□ Show as tooltips

Show structures ? (3)
type to search
☑ s
□ ER_CAUSE
☑ ER_TYPE
☑ TAG
□ EVAL
□ IS
□ CS
□ orig
□ text
□ doc

Show more context ? (off)
Show line selection (on)
Show line copy icon (on)
Show line numbers ? (on)
Use glue ? (on)

Fig 5. Setting the visibility of XML attribute-values in DICO-JALF's SE (Generated by Concordance tool)



CQL <ER_TYPE TAG=="VERB">[] • 266
10,802.91 per million tokens • 1.1%

```
1  □ doc#0皆 は 趣味 が</s><s><ER_TYPE TAG=VERB>  あり  す ね</ER_TYPE></s><s>私 も</s><s><E
2  □ doc#0私 の 趣味 は</s><s><ER_TYPE TAG=VERB>  泳ぎ  ます 。</ER_TYPE></s><s>30メートル く
3  □ doc#0 ></s><s>NA</s><s><ER_TYPE TAG=VERB> 歌おう </ER_TYPE></s><s>多雨</s><s><ER_TYF
4  □ doc#0私 の 趣味 は</s><s><ER_TYPE TAG=VERB>  絵  を かき ます 。</ER_TYPE></s><s>花 を
5  □ doc#0 イタリア で</s><s><ER_TYPE TAG=VERB>  作り  ます 。</ER_TYPE></s><s>この 車 は と
6  □ doc#0;ーゲーム を</s><s><ER_TYPE TAG=VERB> するが 好き です ・</ER_TYPE></s><s>ホラーゲ
7  □ doc#0'ィア など が</s><s><ER_TYPE TAG=VERB>  効き  ます 。</ER_TYPE></s><s>音楽 は 素晴ら
8  □ doc#0 おんがく を</s><s><ER_TYPE TAG=VERB>  きき  ました 。</ER_TYPE></s><s>クラス で
9  □ doc#0 ラマ など が</s><s><ER_TYPE TAG=VERB>  見  ます 。</ER_TYPE></s><s>アニメ が とて
10 □ doc#0 が たくさん</s><s><ER_TYPE TAG=VERB>  あり  ます 。</ER_TYPE></s><s>見る とき 、<
```

Fig 6. Concordance lines from <ER_TYPE> with visible error type in DICO-JALF's SE (Generated by Concordance tool)

concordances. Compare this with Figure 4 above, in which no attribute or value is present.

Objective 2

This subsection demonstrates that the second aim of this present study has also been fulfilled: to measure the accuracy of ChatGPT annotation. It is argued that ChatGPT can annotate errors in approximately 70.31% of the cases. Using a human evaluation procedure, it can be established that ChatGPT is not completely accurate and makes mistakes. See (4). ChatGPT marked できます *dekimasu* 'able to' as an erroneous segment, but this is not accurate as this segment should not be marked as erroneous. The use of dekimasu is indeed correct. Regardless, it was marked as an erroneous segment by ChatGPT. Therefore, the human evaluation is FALSE.

(4) 私は描くことが<IS eval_hum="FALSE">できます<IS>
Watashi-wa-kaku koto-ga-dekimasu
1-TOP-draw-SUBJ-able
'I can draw'

Note that, in many cases, ChatGPT's analyses were accurate. For instance, in (5), it correctly marked 続きます *tsuzukimasu* 'to resume' as an erroneous segment. It also successfully identified VERB_INFL and FORM_ERROR as a correct error type and with correct error cause labels. The correct form of the verb should be 続いています *tsuzuite imasu* 'to resume'. Hence, human evaluators marked them as TRUE, as all evaluations are accurate.

(5) この趣味は子供の時から今まで<IS eval_hum="TRUE"><ER_TYPE TAG="VERB_INFL" eval_hum="TRUE"><ER_CAUSE TAG="FORM_ERROR" eval_hum="TRUE">続きます</ER_TYPE></ER_CAUSE></IS>。
この趣味は子供の時から今まで続きます。
kono-shumi-wa-kodomo no toki-kara-ima-made-

tsuzukimasu
DEM-hobby-TOP-childhood-from-now-until-continue
‘(I) resume this childhood hobby up to now.’

Table 7 shows that ChatGPT was 60.44% accurate in error identification. In terms of error types, accuracy was measured at 65.69%. This means that more than 30% of the error labels were incorrectly assigned. For instance, in (6), ChatGPT assigned ADJ_INFL. However, the correct label should be SPELL, because the erroneous segment is missing い i at the end.

Table 7. Accuracy for ChatGPT performance

No	Segment	Accuracy (%)
1	Correct segment	84.76
2	Erroneous segment	60.44
3	Error type tags	65.69
3	Error cause tags	70.33
4	Mean value accuracy	**70.31**

(6)あのマンガは本当に<ER_TYPE tag=”ADJ_INFL” eval_hum=”FALSE”>かっこい

</ER_TYPE>です。

あのマンガは本当にかっこいです。
Ano-manga-wa-hontou ni-kakkoi-desu
DEM-comic-TOP-really-cool-COP
‘That comic is really cool’

In terms of error-cause labels, the accuracy is slightly better at 70.33 %. This means that only less than 30% of the data were misanalysed (see (7)). In this example, セリス *serisu* ‘serial’ was categorised by ChatGPT as a WRONG_CHOICE in terms of error cause. This is, however, inaccurate because the inaccuracy lies in its form. セリス should be replaced by シリーズ *shiriizu* “serial”, as the former is not listed as valid Japanese vocabulary. The correct label is, therefore, FORM_ERROR, instead of WRONG_CHOICE.

(7)私の趣味はドラマ、アニメ、映画<ER_CAUSE tag=”WRONG_CHOICE”

eval_hum=”FALSE”> セリス</ER_CAUSE>を見ることです。
Watashi-no–shumi-wa-dorama,-anime,-eiga,- serisu-o-miru koto-desu
1-GEN–hobby-TOP-drama-anime-film- serial-OBJ-watch-COP
‘My hobby is watching dramas, animes, movies, and series’

It may be concluded that, in this project, ChatGPT performed with 70.31 % accuracy, averaging across the four parameters (SD= 9.07 Mdn = 68.01, IQR=14.48). This is low compared to the findings of Yu et al. (2024), whose accuracy rate could exceed 90%. One possible reason is that more training data for ChatGPT is available in English instead of Japanese. This data paucity leads to worse system performance when errors in English are identified.

As for the likelihood of error-type labels being valid, an operational question would be: ‘If ChatGPT says this is a particle (PART) error, how accurate is this analysis?’ It can be observed that the error rate for PART (particle) error type labels is the lowest (23.26%), as shown in Table 8. This means the chance of this label being correct is 76.74%. That means the less the error likelihood, the more chances for a label to be correctly annotated.

Table 8. Likelihood of error type labels given by ChatGPT

Error type label	Error likelihood (%)
PRON	84.21
AUX	83.58
WO	78.95
ADJ	71.7
ADV	70.83
VERB	66.92
CONJ	68.18
OTHER	65.52
VERB(INFL)	51.92
NOUN	51.55
ADJ(INFL)	48.65
VERB(TENSE)	46.67
SPELL	39.08
PUNCT	30.43
DET	28.57
PART	23.26

As for the accurate annotations, i.e., all ChatGPT error annotations validated as TRUE by human evaluators, as shown in Figure 7, two of the top-three error tags are related to verbs (VERB and VERB_INFL) with a relatively larger proportion than other errors. This result shows that the acquisition of Japanese verbs for the learners of Japanese in the sample is an area where improvements may be prioritised. Looking closely at the error cause of these categories (all verbs), as shown in Figure 8, the reasons for the verb errors are the use of wrong forms (58%) or the use of incorrect verb choice (38%). Verb overuse accounts for only 21%. This result means that the

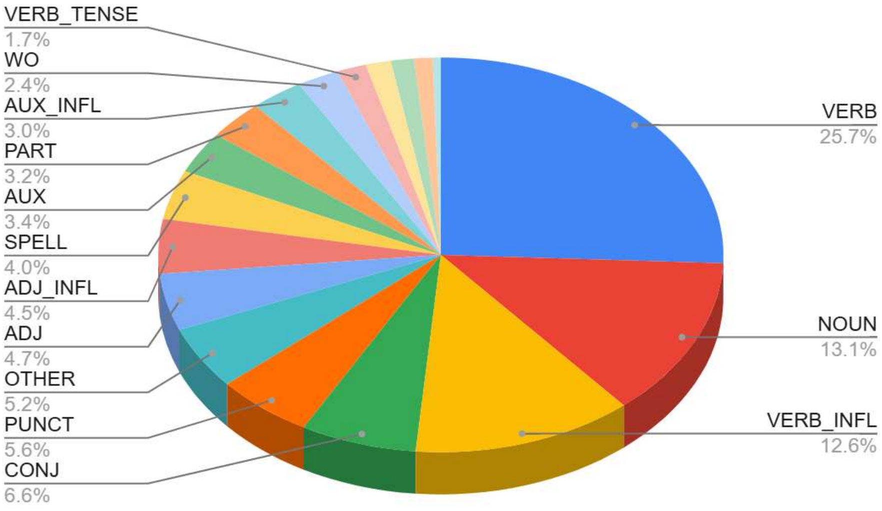


Fig 7. Distribution of error tags

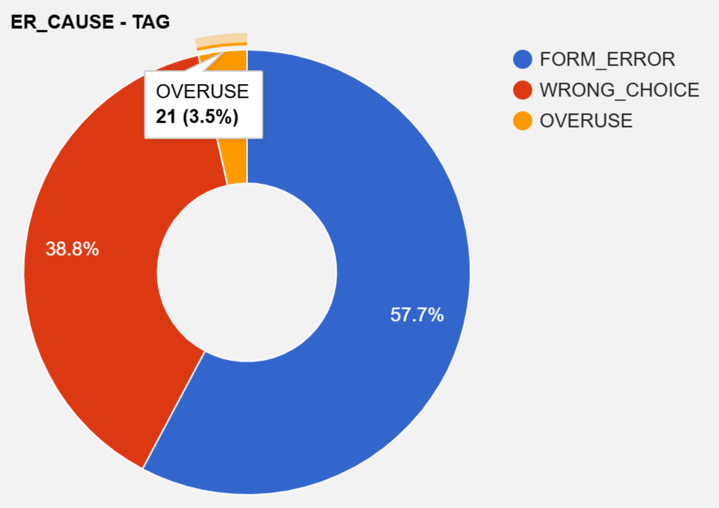


Fig 8. Distribution of error tags for all verbs errors

acquisition of Japanese morphology and semantics for verbs still needs to be improved. These finding aligns with Hayashishita and Ueyama (2020), Yusuf et al. (2024), Putri (2020), and Reina and Lee (2023), who studied verbal errors committed by the learners of Japanese as a foreign language.

In some cases, ChatGPT still assigned tags beyond the required tagset. For instance, it marked the cause of an error as 'PUNCTUATION_ERROR' or 'SPELLING_ERROR'. While the prompts to ChatGPT clearly instructed it not to include categories beyond the required tagsets, this still happened. This may be driven by ChatGPT's temperature. As argued by Poole and Coss (2024), the temperature set (between 0–1) may affect ChatGPT's 'creativity'. As for this research, the default ChatGPT temperature is preserved to experiment how regular users would ChatGPT, which does not require certain technical knowledge. Advanced users can usually implement temperature changes by accessing ChatGPT's API.

IV. DISCUSSION

The results of the study presented earlier can be interpreted as follows. First, regarding error rates, it can be observed that 'verb' is the most problematic area. This, whilst an empirical finding, is anticipated as Japanese has a complex verb conjugation system (Hayashishita and Ueyama, 2020), a feature absent in Indonesian, the L1 of all the respondents in this study. This is substantiated by form errors, as a negative transfer, whose error frequency is quite large. Another substantial cause of mistakes is lexical choice. The difference in the verb semantics arguably causes this. For instance, as shown in Aror et al (2024), the verb 'to wear (outfit)' in Japanese may be realised in different lexical choices depending on the position (as in head, hand, body, feet). And due to the absence of the corresponding lexical choices in Indonesian, they resorted to the wrong verb choice. In terms of ChatGPT accuracy (70% of correct identification of errors), this may be driven by the complexity of the annotation scheme adhered in this study. In a similar study discussing ChatGPT's accuracy (Yu et al., 2024), it may be noticed that the annotation scheme used is different from ours. Another possible factor is ChatGPT 'temperature' (the extent to which ChatGPT can be creative or deterministic) can be applied as shown in Poole and Coss (2024).

The findings show how the gap in corpus-based research on error analysis of Japanese as a foreign language in Indonesia has been filled. Previous studies whose subjects share similarities in their characteristics with ours (Aror et al., 2024; Putri, 2020; Barus and Pujiono, 2021) are not corpus-based. In terms of error annotation scheme and methodology, this study offers some advancements. Most studies study particular errors, while overall error analysis in this study is implemented. Methodologically, previous studies or error analyses of Japanese learners in Indonesia did not use AI tools to identify errors. Instead, errors were directly identified by human teachers. While methodologically different from Sanosi (2022) and Heintz et al. (2022), the study shares similar findings in that, in terms of accuracy, there are some areas in which AI-based applications can still be improved. Regardless of its shortcomings, as shown by Tyson (2023) in the case of ChatGPT, AI is still arguably helpful in providing grammatical feedback. However, it still needs to be overseen by human evaluation to improve its accuracy.

Regarding implications, DICO-JALF v.1.0 (the created corpus) may serve as an empirical database for interlanguage studies. For language pedagogy, the findings may be used as a stepping stone to improve Japanese teaching materials or techniques in, but not limited to, Indonesia. It can also be used for data-driven learning by supplying authentic errors and their corrections. For Natural Language Processing (NLP), the corpus may be used as training data for developing error correction systems.

For future studies, it is also possible to compare ChatGPT with another LLM-based system, such as DeepSeek or Gemini, to compare their performance. In terms of population, the learner group can also be expanded to represent Japanese learners as a foreign language in Indonesia by recruiting subjects from different universities and regions in Indonesian.

V. CONCLUSION

As shown in the earlier section, DICO-JALF v.1.0 has been created, and it is now publicly available (objective 1). This is the first error-annotated learner corpus of Japanese obtained from Indonesian students. This corpus can already be used to support open and distributed learning. Students also have the option to use their preferred systems to access it. The representativeness of the corpus may be improved by incorproating more data from diverse universities in Indonesia.

The use of AI to create error annotations in DICO-JALF has been demonstrated. However, the post-annotation evaluation suggests that the overall annotation accuracy (70%), i.e. the proportion of ChatGPT correctly annotates errors, can still be improved (objective 2). Using this, Japanese language teachers and lecturers can identify errors in different metadata categories and devise data-driven teaching materials, allowing them to target specific areas of weakness.

This research is an essential contribution to AI-based Japanese language assessment tools. In this paper, AI assistance into DICO-JALF to automatically annotate errors has been implemented. In addition, the capability of automated systems to detect and classify common errors in Japanese made by Indonesian students has been demonstrated. This may contribute to the better performance of AI-based language tools.

ACKNOWLEDGEMENT

This research was funded by the Faculty of Humanities, Universitas Diponegoro, through the 2024 Research Grant, International Joint Research Scheme (No: 106/UN7.F6/HK/III), with Prihantoro, S.S., M.A., Ph.D., as the principal investigator, and with Dr. Tanjun Liu (Xi'an Jiaotong–Liverpool University, China) and Prof. Shin'ichiro Ishikawa (Kobe University, Japan) as international research partners. The authors express their sincere gratitude to the reviewers for their insightful feedback and constructive comments, which have greatly contributed to the refinement of this work. Appreciation is also extended to the editors for their professional support throughout the submission and publication process. The authors further acknowledge the invaluable assistance of the research assistants, Ninik Elika, S.S., M.Li., and Haqi Sang Kautsar, S.Li., M.Li. (Cand.), for their dedicated support in data collection and processing.

ETHICS STATEMENT

All respondents whose texts are in the corpus have agreed to make their data publicly available with anonymisation. The informed consent forms for these agreements were digitally signed.

CREDIT AUTHOR STATEMENT

Prihantoro: Conceptualisation, literature review, methodology, analysis, proofreading, writing original draft, writing -- review and editing, visualisation, supervision, project administration, funding acquisition.

Shin'ichiro Ishikawa: Original draft review, methodology, analysis, Japanese native speaker consultant.

Tanjun Liu: Software, statistical analysis, methodology, funding acquisition.

Elizabeth Ika Hesti Aprilia Nindia Rini: Resources, investigation, validation, analysis.

Zaki Ainul Fadli: Resources, investigation, validation, analysis.

Catur Kepirianto: Methodology, analysis, data curation

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Alrehili, A., & Alhothali, A. (2025). Tibyan corpus: Balanced and comprehensive error coverage corpus using ChatGPT for Arabic grammatical error correction. *PeerJ Computer Science*, 11, e2724. https://doi.org/10.7717/peerj-cs.2724

Aror, S., Lensun, S., Sompotan, A., & Pandi, H. (2024). Level of Difficulty Using Verbs Meaning 'To Wear' in Sentences by Students of the Japanese Language Education Study Program. KnE Social Sciences, 322–337. https://doi.org/10.18502/kss.v9i2.14860

Barus, M. B., & Pujiono, M. (2021). An errors analysis in using diathesis (態) in written text of Japanese Language Department senior students at universities in Medan. *IZUMI*, *10*(2), 362–371. https://doi.org/10.14710/izumi.10.2.362-371

Budianto, F. (2023). Welcoming the opportunities: Deciphering contemporary mobility of Indonesian professionals to Japan. *Intermestic Journal of International Studies*, *7*(2), 656. https://doi.org/10.24198/intermestic.v7n2.13

Forti, L. (2023). Learner corpora and the design of data-driven learning activities. In *Proceedings of the EuroCALL 2023 Conference* (pp. 139–144). Polytechnic University of Valencia. https://doi.org/10.4995/EuroCALL2023.2023.16959

Hasibuan, A., & Arfianty, R. (2019). Grammatical and lexical errors of Japanese sentence essay of stikes medistra Lubuk Pakam nurses as apprentices to Japan. *ABDIMAS TALENTA Jurnal Pengabdian Kepada Masyarakat*, *4*(2), 259–266. https://doi.org/10.32734/abdimastalenta.v4i2.4058

Hayashishita, J.-R., Tanaka, D., & Ueyama, A. (2020). A linguistically-informed way of introducing Japanese verbs to second language learners. *Journal of Japanese Linguistics*, *36*(1), 29–72. https://doi.org/10.1515/JJL-2019-2017

Heintz, K., Roh, Y., & Lee, J. (2022). Comparing the accuracy and effectiveness of Wordvice AI Proofreader to two automated editing tools and human editors. *Science Editing*, *9*(1), 37–45. https://doi.org/10.6087/kcse.261

Holland, B. J. (2023). ChatGPT 3.5 and 4. In *Advances in library and information science (ALIS) book series* (pp. 316–340). https://doi.org/10.4018/978-1-6684-7693-2.ch016

Imamura, K., Saito, K., Sadamitsu, K., & Nishikawa, H. (2012). Grammar error correction using pseudo-

Error sentences and domain adaptation. *Meeting of the Association for Computational Linguistics*, *2*, 388–392. https://aclanthology.org/P12-2076/

Imamura, K., Saito, K., Sadamitsu, K., & Nishikawa, H. (2014). Particle error correction from small error data for Japanese learners. *Journal of Natural Language Processing*, *21*(4), 941–963. https://doi.org/10.5715/jnlp.21.941

Park, J. (2019). An AI-based English grammar checker vs. human raters in evaluating EFL learners' writing. *Multimedia Assisted Language Learning, 22*(1), 112–131. https://doi.org/10.15702/mall.2019.22.1.112

Karataş, F., Abedi, F. Y., Gunyel, F. O., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12574-6

Kasahara, S., Komachi, M., Nagata, M., & Matsumoto, Y. (2011). Error correcting Romaji-Kana conversion for Japanese language education. *Workshop on Advances in Text Input Methods*, 38–42. https://aclanthology.info/papers/W11-3506/w11-3506

Koyama, A., Kiyuna, T., Kobayashi, K., Arai, M., Mita, M., Oka, T., & Komachi, M. (2023). Construction of an Error-Tagged Evaluation Corpus for Japanese Grammatical error correction. *Journal of Natural Language Processing*, *30*(2), 330–371. https://doi.org/10.5715/jnlp.30.330

Massey, P. A., Montgomery, C., & Zhang, A. S. (2023). Comparison of ChatGPT–3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *Journal of the American Academy of Orthopaedic Surgeons*. https://doi.org/10.5435/jaaos-d-23-00396

Nishina, K., Hodošček, B., Yagi, Y., & Abekawa, T. (2014). Construction of a learner corpus for Japanese language learners: Natane and Nutmeg. *Acta Linguistica Asiatica*, *4*(2), 37–51. https://doi.org/10.4312/ala.4.2.37-51

OpenAI. (2024). ChatGPT (July-November 2024 version) [Large language model]. https://chat.openai.com/

Pan, M., Guo, K., & Lai, C. (2024). Using artificial intelligence chatbots to support English-as-a-foreign language students' self-regulated reading. *RELC Journal*. https://doi.org/10.1177/00336882241264030

Pavlovič, M. (2020). Grammar errors by Slovenian learners of Japanese: Corpus analysis of writings on beginner and intermediate levels. *Acta Linguistica Asiatica*, *10*(1), 87–104. https://doi.org/10.4312/ala.10.1.87-104

Pérez-Paredes, P., & Mark, G. (2022). What can corpora tell us about language learning? In *Routledge eBooks* (pp. 313–327). https://doi.org/10.4324/9780367076399-22

Poole, F. J., & Coss, M. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). *Journal of Technology for Chinese Language Teaching, 15*(1), 1–24. http://www.tclt.us/journal/2024v15n1/poolecoss.pdf

Pandey, A. (2022). Improving Hindi POS Tagger Accuracy Through Domain Adaptation. Nepalese Linguistics, 79–85. https://doi.org/10.3126/nl.v35i01.46564

Prihantoro. (2022). SANTI-morf Dictionaries. *Lexicography, 9(2),* 175-193. https://doi.org/10.1558/lexi.23569

Prihantoro. (2025). The creation of the Indonesian TreeTagger for use in LancsBox and CQPweb (2). 13(2). https://doi.org/10.32714/ricl.13.01.11

Putri, M. A. (2020). Morphological errors on Japanese verb conjugation to passive form at third-year students of Japanese education study program of UNP. *Proceedings of the Eighth International Conference on Languages and Arts*, *258-263*. Negeri Padang, Indonesia: Atlantis Press. https://doi.org/10.2991/assehr.k.200819.052

Reina, K., & Lee, J. (2023). An Analysis of Verb Errors Made by Japanese Learners of Korean. *Studies in Linguistics*, *67*, 199–220. https://doi.org/10.17002/sil..67.202304.199

Safama, S. A., & Diner, L. (2022). Analisis kesulitan penggunaan partikel wa, no, ni, de pada siswa MAN 1 Kebumen. *JLA (Jurnal Lingua Applicata)*, *6*(1), 44. https://doi.org/10.22146/jla.75070

Sanosi, A. (2022). To err is human: comparing human and automated corrective feedback. *Information Technologies and Learning Tools*, *90*(4), 149–161. https://doi.org/10.33407/itlt.v90i4.4980

Thewissen, J. (2013). Capturing L2 Accuracy Developmental Patterns: Insights From an Error-Tagged EFL Learner Corpus. The Modern Language Journal, 97(S1), Article S1. https://doi.org/10.1111/j.1540-4781.2012.01422.x

Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, *100*(8), 3098–3101. https://doi.org/10.1021/acs.jchemed.3c00361

Umoro, A. L. (2023). Indonesian students' motivation to pursue tertiary education in Japan. *Japanese Research on Linguistics Literature and Culture*, *5*(2), 118–127. https://doi.org/10.33633/jr.v5i2.8282

Yang, J. C., & Akahori, K. (2013). Error analysis in Japanese writing and its implementation in a computer assisted language learning system on the world wide web. *CALICO Journal*, *15*(1–3), 47–66. https://doi.org/10.1558/cj.v15i1-3.47-66

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.23087.yu

Yusuf, M., Taulia, T., & Mawaddah, A. (2024). An Analysis of Japanese Verb Conjugation Errors Among Students at Universitas Harapan Medan. *Humanities & Language*, *1*(3), 183–190. https://doi.org/10.32734/w2gj8r54