GRAMMAR ERRORS BY SLOVENIAN LEARNERS OF JAPANESE: CORPUS ANALYSIS OF WRITINGS ON BEGINNER AND INTERMEDIATE LEVELS

Miha PAVLOVIČ
University of Ljubljana, Slovenia
miha.pavlovic1@gmail.com

Abstract

This paper presents the construction of a corpus of writings by Slovene learners of Japanese as a foreign language at the beginner and intermediate levels and an analysis of the grammar errors contained within it, with the purpose of providing a simple and effective means of acquiring data on errors made by students of Japanese as a second language. Additionally, an error analysis of the grammar errors in the corpus and a comparison of the most common errors found on both levels, reveals the types of errors that carry over from the beginner to the intermediate level, negatively affecting the learning process. By compiling and analyzing a collection of 182 written texts written by Japanese learners, 492 cases of grammar misuse were observed on the beginner and 564 on the intermediate level. A comparative analysis of the most common types of grammar misuse on each level highlights the types of errors that seem to carry over from the beginner to the intermediate level. The findings can be useful to Japanese language learners as well as teachers. Furthermore, the learner's corpus created in the process marks the first step towards the creation of a larger, annotated and publicly accessible learner corpus of writings by Slovenian learners of Japanese to be used for further research in the field of second language acquisition.

Keywords: learner corpus; corpus construction; error analysis; grammar error; second language acquisition

Povzetek

Članek opisuje izgradnjo korpusa usvajanja jezika slovenskih študentov japonščine na osnovni in srednji ravni in analizo slovničnih napak v njem z namenom ustvarjenja orodja, ki bo uporabnikom omogočalo na enostaven in pregleden način pridobiti podatke o najpogostejših napakah v spisih slovenskih učencev japonščine in s pomočjo analize napak v le-tem ugotoviti, katere slovnične strukture povzročajo največ težav slovenskim učencem japonskega jezika na posamezni ravni ter s pomočjo primerjave rezultatov izpostaviti tipe napak, ki se prenašajo iz osnovne na srednjo raven. Korpus vsebuje 182 spisov, v katerih so označene in kategorizirane napake. Napak je 492 na osnovni in 564 na srednji ravni. S primerjavo najpogostejših napak na posamezni ravni so se bili izpostavljeni tipi napak, ki se prenašajo iz osnovne na srednjo raven. Te ugotovitve lahko koristijo tako učencem kot tudi učiteljem japonščine pri učnem procesu, hkrati pa je tako nastali korpus prvi korak k izgradnji obsežnega, označenega in javno dostopnega korpusa besedil slovenskih učencev japonščine za nadaljnje raziskave o učenju japonščine kot tujega jezika.

Ključne besede: korpus usvajanja jezika; gradnja korpusa; analiza napak; slovnične napake; usvajanje tujega jezika

Acta Linguistica Asiatica, 10(1), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.1.87-104

1 Introduction

The Slovenian and Japanese language are genealogically not related and thus differ on all levels of linguistic analysis: from script and phonology to grammar and syntax. At the syntactic level, the predicate in Slovene sentences mostly appears in second place, usually following a subject or adverbial, while in Japanese the predicate always appears at the end of a sentence or subordinate clause. On the grammatical level, there is a difference in the way cases are expressed; while in Slovene cases are expressed by noun declension, in Japanese particles (*kakujoshi* 格助詞) are attached to grammatical elements to mark their relation to the verb; while Japanese adjectives ending with an -i (*i-keiyōshi* イ形容詞) have past forms, Slovenian adjectives do not have different forms to express tense and a past form of the auxiliary verb is used, and there are many other subtler differences. It is therefore considerably more challenging and time-consuming for a Slovenian learner to learn Japanese than a more related language like English or German, which share grammatical similarities with the Slovenian language.

The occurrence of grammar errors is a natural part of the language acquisition process; thus it is only natural that learners make more errors when using the elements that are fundamentally different from those in their native language. The reason for the occurrence of such errors is usually attributed to the lack of knowledge about those elements. If such errors can be recognized and corrected, a strong foundation for further language acquisition may be guaranteed. Some types of errors disappear naturally, through exposure to the language. However, some errors, if not recognized and dealt with, persist and negatively influence the process of language acquisition. For these purposes, researchers in the field of second language acquisition (SLA) conduct so called "error analyses", which, as the name suggests concern themselves with the quantitative and qualitative analysis of the errors produced by learners of a specific language. The tools used in such studies most commonly include databases or corpora containing examples of language use by students of a specific skill level (e.g. English learners on the intermediate level).

Due to the field of Japanese studies in Slovenia being fairly new, similar studies focused on the errors made by Slovenian learners of Japanese have been very few in number. Thus, there was a lack of and need for a tool that would allow users to easily access data on the types of errors Slovenian learners of Japanese tend to make in written compositions on a certain level. One of the aims of the present study is therefore, through the acquisition and digitalization of learners' compositions, to create a corpus of errors by Slovenian learners of Japanese on both the beginner and intermediate level. Grammar errors on both levels were analyzed with the purpose of exposing the problematic grammatical elements that are prevalent on both levels. As mentioned previously, such types of errors, when unidentified, may hinder language acquisition. Expozing and consequently targeting them can have a positive effect on the learning process.

In short, the purpose of the study was to produce a resource in which teachers and SLA researchers can easily access data on the types of grammar errors Slovenian learners of Japanese tend to make, and by using the data expose the most problematic grammar error types.

Sections 2 contains a summary of previous research, used as reference. Sections 3 to 5 describe the creation of the corpus: section 3 the metadata added to the students' compositions, section 4 the process of data acquisition and digitalization, and section 5 the categorization of error types. The second part of the paper presents a first analysis of this corpus: section 6 describes the methodology used in the analysis, sections 7 and 8 the results of the analysis of grammar errors on the beginner and intermediate levels respectively, section 9 a comparison of the results on each level, followed by their discussion in section 10 and conclusions in section 11.

2 Previous research on errors in a second language acquisition

In the last decades, a number of error analyses targeting the grammar errors made by foreign students of Japanese (mostly native speakers of English, Chinese and Korean) have been conducted, mostly by Japanese linguists. Examples of such studies include: Teramura (1990), Ichikawa (1993), Kawaguchi (1995), Otsuka & Hayashi (2010), Harasawa (2012), Noda and Sakoda (2019) and others.

Present research is the first study to analyze a corpus of Slovenian Japanese learners, and as such seeks to verify whether the findings from previous studies are valid for native speakers of Slovene as well.

The following three surveys were primarily used as an important source of information and guidance for this analysis.

Kawaguchi (1995) analyzed writings of five students with different middle-level native languages. The compositions averaged around 400 characters, which caused 267 cases of errors. The most numerous types of errors involved particles, case particles in particular. The author concluded that such types of errors are often carried over to the advanced level. The comparison of the results for different levels of acquisition was taken as a model for the present research.

Han 2014 identified 2875 errors using quantitative analysis of 204 compositions. Grammar and semantic errors together accounted for almost 90 % of all errors, of which grammatical presented as much as 54.6 % while 33.8 % were semantic. The most common type of grammar errors (30.5% of all grammar errors) associated with the group of articles, of which case particles were found to be most problematic and represented 65% of all errors related to the use of particles. The most common mistakes were made in distinguishing between the use of *ga* が and *wa* は. Similar difficulties was also observed with the distinctions between: *ni* に, *de* で, *wo* を, *ga* が and no の. Methodology used in this research was a model for our research.

Finally, Online Dictionary of Errors in Japanese 2011 (*Onrain nihongo goyō jiten* オンライン日本語誤用辞典 2011), created at the University of Foreign Languages in Tokyo, is introduced not only for its error analysis but also as a tool designed to further conduct this type of research. The tool is based on a corpus containing more than 1000 entries of errors identified from 40 files, totaling more than 20,000 characters. The online dictionary is currently one of the few, if not the only, online corpora or dictionary that categorizes collected errors on multiple levels and allows the user to view them in a simple and transparent way. This online glossary is very important for the present research because the categorization used in building the corpus is based on the categorization of errors used in this corpus.

3 Slovenian learners of Japanese: corpus analysis of grammar errors

3.1 Methodology

3.1.1 Metadata structure and annotation

The Slovenian learners' written Japanese corpus consists of two sub-corpora: *Slovenian beginner learners' written Japanese corpus* and the *Slovenian intermediate learners' written Japanese corpus*.

The sub corpus of the beginner level consists of 142 shorter compositions, each with an average length of about 280 characters. The compositions were written by 29 first-year students of the Japanese studies program at the Department of Asian Studies in the Faculty of Arts, University of Ljubljana in the academic year 2016/2017. The compositions were not written in a test environment, but as homework at two of the Japanese language classes. The topics of the compositions cover a range of simple everyday topics (9 in total), such as descriptions of one’s room, one’s family, hobbies, a diary, a self-presentation and a reading diary.

The sun-corpus of the intermediate level consists of 40 longer compositions, each with an average length of about 500 characters. The compositions were written in 2017/2018 by 11 of the same 29 students (one year later than the first compositions). The compositions include 4 topics which require the use of more complex grammatical structures and vocabulary than the topics of the beginner corpus, and the students were asked to state and argue their opinion on the subject. These topics are: “telephone”, “time”, “world heritage” and “my country”. These compositions were written as part of a mid-term exam, where dictionaries and grammar checkers were not allowed.

3.1.2 Acquisition and digitalization of the compositions

The compositions were submitted as homework or parts of mid-term exams. Each of the authors signed a waiver, allowing the inclusion of their compositions into the corpus and their use scientific purposes, under the condition that all personal data be anonymized.

The next step was digitization. The creation of the corpus required a tool for the annotation of grammar errors and search of both specific parts of the data (compositions), as well as the metadata (categories, data on the compositions, etc.), easy acquisition of statistical data and that would be portable on and compatible with different platforms (Mac, Windows, etc.). While several sets of open-source annotation software (such as “Slate”, “WebAnno”, “SketchEngine” and others) were available, none of these tools appeared to satisfy all of the required criteria. The tool that finally provided an almost surprisingly simple solution to the problem was Microsoft’s Excel.

First, all of the texts were manually typed into a Microsoft Excel spreadsheet (each sentence in a separate row) verbatim as they appeared in the handwritten physical version; all errors, including orthographical errors, errors in the use of *kanji* 漢字, were transcribed as in the original.

This was done to enable the created corpus to be used for different types of error analysis in the future and to provide possible context for the occurrence of errors. All personal data was anonymized and replaced with a placeholder (*jinmei* [人名] for personal names, or *chōmei* [町名] in the case of town names).

Non-standard character forms were not annotated, because the inclusion of such errors would require a fairly different approach and toolset. Thus it seemed best to omit these types of errors.

The final step of the digitization process was error annotation. All error annotation from the original correction, done by the teacher in charge of the class, was carried over. Where annotations other than those made by the teacher were marked differently from the original annotations.

Finally, in a separate spreadsheet, a corrected version of each sentence containing an error was added in a column next to the original sentence and the corrected part marked with one of three colors, depending on the type of error: red for grammar errors, yellow for orthographical errors or errors connected to the use of Chinese characters and green for stylistic errors and errors in vocabulary choice.

3.1.3 Error categorization

While other types of errors were also included and annotated in the corpus, the analysis described in this paper focuses solely on grammar errors. Each of the grammar errors was categorized first into a main group, followed by a subgroup and finally within each subgroup according to the supposed cause for the error. However, when being categorized, the error was not categorized according to the grammar element that was mistakenly used, but according to the element that should have been used to form a grammatically correct sentence. The basis for this is the idea that, as mentioned in the first chapter, the cause for the occurrence of the error is a lack of knowledge about the element; in this case knowledge of the fact that this specific element needed.

Table 1: Examples of grammar errors due to a wrong choice

Grammatically incorrect sentence	Sentence as corrected by teacher	Error	Grammatical category	Sub-category	Cause of Error
ゲームを好きです。	ゲームが好きです。	が ⇔ を	格助詞	が	誤選択
Gēmu wo suki desu.	*Gēmu ga suki desu.*	*Ga ⇔ wo*	*kakujoshi*	*ga*	*gosentaku*
I like games.	I like games.		Case particle	Particle *ga*	Wrong choice

As seen in the above table, in the sentence "*Gēmu wo suki desu.*" the grammar error occurred due to the student using the particle *wo* instead of the particle *ga*, which this sentence structure calls for. The error would be classified as an error connected to the use of case particles, more precisely, the case particle *ga*, with the contributing cause being marked down as *wrong choice*.

Table 2: Examples of grammar errors due to lack of use

Grammatically incorrect sentence	Sentence as corrected by teacher	Error	Grammatical category	Sub-category	Cause of Error
ゲーム Ø 好きです。	ゲームが好きです。	が ⇔ Ø	格助詞	が	誤不足
Gēmu suki desu.	*Gēmu ga suki desu.*	*Ga ⇔ Ø*	*kakujoshi*	*ga*	*gofusoku*
I like games.	I like games.		Case particle	Particle *ga*	Lack of use

In the case of the sentence "*Gēmu suki desu.*" the grammar error occurred due to the student not using the particle *ga*; therefore, this type of error would again be categorized as an error connected to the use of the case particle *ga*, the difference here being that the contributing cause would be marked as "lack of use".

This categorization was adopted from the categorization used in a similar learner's corpus of Japanese learners' grammar errors, namely the *Online corpus of Japanese* *learners' errors* by Umino's et al. (2012, originally: *Onrain nihongo goyō jiten* オンライン日本語誤用辞典) published by the Tokyo University of Foreign Studies.

The reason for this choice is that the former corpus is one of the few corpora of Japanese learners in which errors are not only annotated, but also categorized in groups and subgroups according to their grammatical properties in a very similar manner as demonstrated in the above table. The reason an already existent classification was used was to make the data in these two corpora easily comparable, thus further increasing the number of possible uses for the assembled data in potential future studies.

Following below are three tables. The first contains all the main grammatical categories used. The second one contains the sub-categories of specific types of elements within each of the main grammatical categories. And the third table contains the five types of contributing causes that were determined for each error. The left column of each table contains the Japanese name of the category accompanied by its transcription and the right one an English translation by the author.

Table 3: Grammatical categories

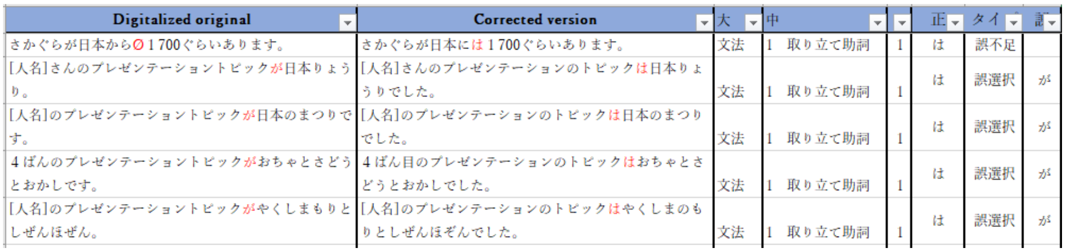
	Japanese original	Transcription	English translation
1-1	取り立て助詞	*toritatejoshi*	focus particles
1-2	格助詞	*kakujoshi*	case particles
1-3	終助詞	*shūjoshi*	final particles
1-4	複合辞	*fukugōji*	compound particles
1-5	ヴォイス	*voisu*	voice
1-6	テンス・アスペクト	*tensu-asupekuto*	tense and aspect
1-7	基本文型	*kihonbunkei*	basic sentence structure
1-8	表現文型	*hyōgenbunkei*	modal expressions
1-9	待遇表現	*taigūhyōgen*	polite expressions
1-10	形式名詞	*keishikimeishi*	formal nouns
1-11	指示詞	*shijishi*	demonstratives
1-12	疑問詞	*gimonshi*	interrogatives
1-13	2語の接続	*ni-go no setsuzoku*	word level conjunction
1-14	2文の接続	*ni-bun no setsuzoku*	sentence level conjunction
1-15	修飾	*shūshoku*	modifiers

Table 4: Error causes

Japanese original	*Transcription*	English translation
誤選択	*gosentaku*	wrong choice
誤不足	*gofusoku*	lack of use
誤形態	*gokeitai*	form error
誤付加	*gofuka*	redundance
誤位置	*goichi*	wrong position

In order to classify and annotate the errors, a framework needed to be created, so as to create space for the marks, enabling the different functions of MS Excel to work as intended.

As mentioned in the above paragraphs, the original text was placed in an excel spreadsheet, accompanied by the corrected version in the neighboring column. The column next to it (column C in the example bellow) contains the data on the type of error, ranging from "grammar", "style and vocabulary" to "orthography and script". The fourth column was created for data on the grammatical category and the one next to it for data on the grammatical sub-category (as explained in 4.1) to be inserted. The sixth column was made for data on the specific grammar element that was supposed to be used in the sentence where the error occurred (in some cases this data was the same as that in the fifth column, however in cases where the sub-category was an umbrella term, such as "temporal conjunctions" it served to further pinpoint the specific type of error). The seventh column was used to determine the cause of the error, while the eight one was used to mark which element was wrongly used instead of the right one. The final, ninth column was used to add numerical IDs to each of the sentences, making it possible to restore their original order within the whole framework after using different sorting options in Excel.



Digitalized original	Corrected version	大	中		正	タイ	誤
さかぐらが日本からØ 1 700ぐらいあります。	さかぐらが日本には 1 700ぐらいあります。	文法	1　取り立て助詞	1	は	誤不足	
[人名]さんのプレゼンテーショントピックが日本りょうり。	[人名]さんのプレゼンテーションのトピックは日本りょうりでした。	文法	1　取り立て助詞	1	は	誤選択	が
[人名]のプレゼンテーショントピックが日本のまつりです。	[人名]のプレゼンテーションのトピックは日本のまつりでした。	文法	1　取り立て助詞	1	は	誤選択	が
4ばんのプレゼンテーショントピックがおちゃとさどうとおかしです。	4ばん目のプレゼンテーションのトピックはおちゃとさどうとおかしでした。	文法	1　取り立て助詞	1	は	誤選択	が
[人名]のプレゼンテーショントピックがやくしまもりとしぜんほぜん。	[人名]のプレゼンテーションのトピックはやくしまのもりとしぜんほぞんでした。	文法	1　取り立て助詞	1	は	誤選択	が

Figure 1: Example of corpus

This design now enables the user to use Excel's sorting and search functions to e.g. search for all the instances of a specific error, to find all the cases in which a specific grammar element was used wrongly, sort the data according to each of the three categories (grammatical category, grammatical sub-category and error cause), search for specific terms used in either the original or the corrected data, easily acquire statistical data for cases of any of the above, and many more.

3.1.4 Data analysis

By using Microsoft Excel's sort and search functions the number of errors correlating to each group was counted for all the categories mentioned in Chapter 5.

The number of errors in each group and sub-group were then compared to the sum of all errors and were henceforth represented with percentages rather than actual numbers. This was also partially done to enable easier comparison of the results on each level in the second part of the analysis.

Next the grammatical categories and sub-categories with the highest amount of errors were determined alongside the most common causes for the occcurance of each type of error.

However, it must be said, that the percentages of errors described in the following sections are not a direct indicator of the relative difficulty of a particular morphological or syntactic category, only of the frequency of errors being made. To determine the relative difficulty of specific categories, a different approach would be necessary.

The number of errors related to categories that are more frequent (e.g. case particles) is necessarily larger than the number of errors related to categories that are less frequent in any text (e.g. final particles).

Originally, one of the goals was to identify the most numerous types of errors, and based on the ratio between the amount of correct and incorrect use of an element. However, because of the relatively small amount of data in each sub corpus and the uneven use of different grammatical elements within it, the calculated results were unreliable. In addition, in previous research, which served as the basis for this analysis, this step was also omitted.

Finally, if a learner were to misuse a rarely used grammar in 10 out of 10 cases, compared to a more common grammar being misused 200 out of 500 cases, the latter type of error would hinder communication between the author and the reader much more, simply because of its frequency. Additionally, the calculation itself would be too time consuming, in proportion to the unreliable results to be gained. Thus, only misuse frequency was determined.

This was done with both the Slovenian beginner learners' written Japanese corpus and the Slovenian intermediate learners' written Japanese corpus respectively. Thus the results on both levels were acquired and the elements the students struggle with the most were determined. The results are presented in the following sections.

The next step was the comparison of the results on both levels. As mentioned in chapter 1, this was done with the goal of exposing the types of grammar errors that appear on the beginner level and are still present on the intermediate level. The persistence of such errors means that they present a huge hurdle to the learner, which, if not overcome, would exert negative influence on the language acquisition process further on.

To expose these errors, the appearance rate of each of the error groups and sub-groups was observed and compared.

As a result, groups of problematic grammatical elements were successfully exposed and analyzed. A detailed summary of the results can be found in the following sections.

3.2 Results

3.2.1 Grammar errors on the beginner level

The beginner sub corpus includes 142 compositions in which 496 grammar errors were observed. The average length of the compositions is about 210 Japanese characters.

Table 5: Error data on the beginner level

Error category	Number of occurrences	Percentage of all errors
1-2 case particles	129	26,2 %
1-7 basic sentence structure	100	20,2 %
1-15 modifiers	74	15,0 %
1-6 tense and aspect	58	11,8 %
1-14 sentence level conjunction	40	8,1 %
1-1 focus particle	38	7,7 %
1-5 voice	12	2,4 %
1-10 formal nouns	12	2,4 %
1-4 composed particles	10	2, 0%
1-8 modal expressions	7	1,4 %
1-13 word level conjunction	6	1,2 %
1-3 final particles	3	0,6 %
1-11 demonstratives	3	0,6 %
1-12 interrogatives	1	0,2 %
SUM	**493**	**100,0 %**

The most common error categories, sorted from most to least common, are presented in table 5. Most errors were found in the group of case particles, amounting to 26 % of all errors found. The second most common were errors connected to the basic sentence structure which represent 20,2 % of all errors found; the third most common being the group of modifiers with 14,9 % of all errors. A considerable number of errors was also found in the category of sentence level conjunctions with a sum of 8,1 % and focus particles with 7,7 %.

Table 6: Error cuases on the beginner level

Type of error	Number of occurrences	Percentage of all errors
wrong choice	214	43,4 %
lack of use	164	33,3 %
form error	55	11,2 %
redundance	50	10,1 %
wrong position	10	2 %
SUM	**493**	**100,0 %**

As represented in the table above, the most common cause of errors was wrong choice with 43,4 %, followed by lack of use with 33,3 % of all cases. Other causes were much less common.

3.2.2 Grammar errors on the intermediate level

The subcorpus of Slovenian intermediate learners' written Japanese contains 40 compositions in which 564 grammar errors were observed. The average length of the compositions amounts to about 550 Japanese characters per composition.

Table 7: Errors on the intermediate level

Error category	Number of occurrences	Percentage of all errors
1-2 case particles	127	22,5 %
1-14 sentence level conjunction	99	17,6 %
1-1 focus particles	95	16,8 %
1-6 tense and aspect	51	9,0 %
1-7 basic sentence structure	42	7,4 %
1-8 modal expressions	36	6,4 %
1-15 modifiers	33	5,9 %
1-5 voice	29	5,1 %
1-10 formal nouns	23	4,0 %

Error category	Number of occurrences	Percentage of all errors
1-11 demonstratives	13	2,3 %
1-4 composed particles	9	1,6 %
1-13 word level conjunction	6	1,0 %
1-3 final particles	1	0,2 %
1-12 interrogatives	0	0,0 %
SUM	**564**	**100,0 %**

As seen in the table above, the most common errors were those related to the use of case particles with 22,5 % of all the errors observed. Also very common were errors from the categories of sentence level conjunction (17,6 %) and focus particles (16,8 %). Errors from the category tense and aspect (9 %), basic sentence structure (7,4 %) and modal expressions (6,4 %) were also common.

Table 8: Error cuases on the intermediate level

Type of error	Number of occurrences	Percentage of all errors
wrong choice	322	57,1 %
lack of use	133	23,6 %
redundance	64	11,3 %
form error	42	7,4 %
wrong position	3	0,5 %
SUM	**493**	**100,0 %**

As can be seen in the above table, the predominantly common cause of errors was wrong choice *gosentaku* with 57,1 %, followed by lack of use *gofusoku* with 23,6 % of all cases. Other causes were much less common.

3.2.3 Comparison of the results on both levels

After grammar analysis on each level was completed, a comparative analysis of the results on both levels was conducted. First, we will present comparison of the most common error categories, which will be followed by comparison of error causes across both levels.

The following table presents a comparison between the most common error categories on each level (as described in chapters 6 and 7). The categories in which a difference of more than 2 % was observed between the beginner and intermediate level are marked with blue if the percentage decreased, and red if the percentage increased. The threshold was first set to 5 %, but was later lowered down to 2 %, to accommodate for and include categories with differences between the two levels lower than than 5 %.

Table 9: Comparison of analysis results on both levels

Analysis of errors on the beginner level			Analysis of errors on the intermediate level		
1-2 case particles	129	26,2 %	1-2 case particles	127	22,5 %
1-7 basic sentence structure	100	20,2 %	1-14 sentence lev. conjunction	99	17,6 %
1-15 modifiers	74	15,0 %	1-1 focus particles	95	16,8 %
1-6 tense and aspect	58	11,8 %	1-6 tense and aspect	51	9,0 %
1-14 sentence lev. conjunction	40	8,1 %	1-7 basic sentence structure	42	7,4 %
1-1 focus particle	38	7,7 %	1-8 modal expressions	36	6,4 %
1-5 voice	12	2,4 %	1-15 modifiers	33	5,9 %
1-10 formal nouns	12	2,4 %	1-5 voice	29	5,1 %
1-4 composed particles	10	2,0 %	1-10 formal nouns	23	4,0 %
1-8 modal expressions	7	1,4 %	1-11 demonstratives	13	2,3 %
1-13 word level conjunction	6	1,2 %	1-4 composed particles	9	1,6 %
1-3 final particles	3	0,6 %	1-13 word level conjunction	6	1,0 %
1-11 demonstratives	3	0,6 %	1-3 final particles	1	0,2 %
1-12 interrogatives	1	0,2 %	1-12 interrogatives	0	0,0 %
SUM	493	100 %	SUM	564	100 %

By comparing the two tables, in 8 of the 14 categories changes in appearance percentage can be observed. At the transition from beginner to intermediate level a decrease of occurrence can be seen in errors connected to the use of:

- case particles (26,2 % → 22,5 %) – however still the most common error category;
- basic sentence structure (20,2 % → 7,4 %);
- modifiers (15 % → 5,9 %);
- tense and aspect (11,8 % → 9,0 %).

An increase in occurrence can be seen in errors connected to the use of:

- sentence level conjunction (8,1 % → 11,8 %);
- focus particles (7,7 % → 16,8 %);
- voice (2,4 % → 5,1 %9;
- modal expressions (1,4 % → 6,4 %).

Aditionally, by comparing the two tables a more equal spread of error percentage across all categories can be observed. This can be explained by the fact that the students on the intermediate level use a wider range of grammatical structures and grammar types from all groups, which causes a higher diversity in error types.

Below is a table comparing the supposed causes attributed to the errors on each level.

Table 10: Comparson of error causes on both levels

Analysis of errors on the beginner level			Analysis of errors on the intermediate level		
wrong choice *gosentaku*	214	43,40 %	wrong choice *gosentaku*	322	57,10 %
lack of use *gofusoku*	164	33,30 %	lack of use *gofusoku*	133	23,60 %
form error *gokeitai*	55	11,20 %	addition *gofuka*	64	11,30 %
addition *gofuka*	50	10,10 %	form error *gokeitai*	42	7,40 %
wrong position *goichi*	10	2 %	wrong position *goichi*	3	0,50 %
SUM	493	100 %	SUM	564	100 %

Through comparison of the results, the following conclusions can be drawn:

- the most common cause of errors on both levels is due to wrong choice;
- at the transition from beginner to intermediate level an increase in the errors caused by wrong choice can be observed;
- on both levels a considerable ammount of errors was also caused by lack of use – however the percentage decreased by almost 10 % when transitioning to the intermediate level;
- the errors caused by error in form decreases when transitioning to the intermediate level.

4 Overall discussion

The following subsections compare the results of the error analysis on the beginner and intermediate level.

4.1 Determining problematic errors

In the cases where a substantial reduction in the appearance rate of an error category was observed, it was interpreted as, depending on the degree of reduction, successfully alleviated; on the other hand, error groups in which a decrease in appearance rate was hardly present, non-existent or an increase of appearance rate was observed, were interpreted to be potentially problematic and were therefore marked and examined more carefully.

4.2 Errors concerning particles

Error types connected to particles (especially case particles) tend to carry over from the beginner level to the intermediate level, and are the most common type of errors on both levels.

Errors in the use of the case particle *ga* tend to carry over to the intermediate level most; while the most common cause for such mistakes is confusing its use with the focus particle *wa*.

Errors connected to the use of the focus particle *wa* present one of the most common error types on both levels. With the transition to the intermediate level an increase of such levels can be observed. This suggests that a further increase might be present in the transition to the advanced level as well. Most commonly the cause of these errors is due to confusing its use with the case particle *ga*.

While errors connected to the case particle *wo* do tend to carry over to the intermediate level, they appear less commonly.

Errors connected to the case particles *de* and *ni* are especially common and seem to carry over to the intermediate level. The predominant cause for these errors is due to learners confusing the use of one with the other.

Errors connected to the attributive particle *no* present the most common type of error on the beginner level. However, through the transition to the intermediate level these types of errors are far less common, which suggests that the learners seem to be growing accustomed to its use. A further decrease might appear at the transition to the advanced level.

4.3 Other error groups

On the beginner level learners seem to struggle with the use of the copula *da*/*desu*. Such errors are hardly present on the intermediate level.

Errors connected to verb and adjective conjugation are very rare on the intermediate level, in contrast to their prevalence on the beginner level, indicating that learners on the intermediate level are already fairly familiar with the conjugations and forms of the adjectives and verbs, thus most of the cases of misuse actually appear to be mistakes rather than errors. The difference between the two is that mistakes happen accidentally (typos, etc.), unlike errors, which happen due to a lack of knowledge (the student has incorrect information on the use of a specific grammatical element).

The same reduction can be observed with errors connected to the use of the past tense of adjectives and verbs.

Errors in the use of sentence level conjunctions are less common on the beginner level, where the learners are only familiar with a small amount of such grammatical structures. They were mostly observed in cases of enumeration and basic sentence conjunctions. On the intermediate level however, an increase in all of the subcategories was observed. This can be attributed to the fact that the learners on the intermediate level are familiar with a much wider range of different conjunctions, which makes for a higher chance of an incorrect one being used. Furthermore, in many cases the errors occur due to conjunctions being mistakenly used in the place of other conjunctions within the same subcategory (i.e. potential clauses).

4.4 Error causes

The types of errors that proved most persistent were those caused by wrong choice – errors where a grammatical element is used instead of another one.

Errors caused due to wrong form of a grammatical element are fairly common on the basic level, but tend to disappear when transitioning to the intermediate level.

4.5 Comparison to previous studies

When comparing the results of the analysis with those of preceding analysis' quite a few similarities can be observed. Similar to Ichikawa (1993) the ratio of errors due to misuse of conjunctions is fairly high. Similar to Kawaguchi (1995) and Yō (2014) the most common type of mistakes are mistakes connected to the use of particles, especially case particles.

5 Conclusions

Having conducted the present research, we have recognized several limits and will here introduce possibilities for their improvement.

The first point we would like to highlight is the scope of the corpus. It is currently comprised of 182 texts (142 shorter and 40 longer) written by students at both levels. Compared to other corpora, this number is quite low. For the purposes of future research, and in particular to increase the credibility of the results, both sub-corpora will need to be expanded and a corpus of advanced learners added.

Another point that should be improved is the categorization. Initially, the categorization was created to be used with a corpus, but given that it was not made specifically for this one, categorization, made specifically for this corpus should be made. Yō 2014 also highlights the lack of a generally established standard for categorizing grammar errors in the Japanese language as a common problem.

Usually, when annotating and categorizing errors in the creation of a learner's corpus, the work is done in groups, then the errors are determined according to the most commonly marked category. Because the categorization process has mostly been done individually a revision of the categorized errors will be needed. When the corpus is made publicly available, a system, that allows the users to submit suggestions or report errors will be set up, so that the corpus and the data within can constantly keep evolving and improving.

Another possibility for improvement is the optimization of software used as a corpus framework. As mentioned in 4.2, Microsoft Excel is currently used for the corpus framework. Although it currently meets all the needs of the corpus and has many positive features, with the growth of the corpus there will also be a need for a tool that makes it easier to add and annotate texts, analyze content and the like.

Last but not least, while findings obtained from both of the sub-corpora analyzes certainly provide useful data with a sufficient degree of credibility, due to the small size of the corpus, an adequate measure of criticality is also required when interpreting the results. As mentioned in the introduction, the purpose of the analysis was to provide students and teachers with an insight into the most common types of grammar errors and to, through the construction of the corpus, take the first step towards the final goal of an online corpus of Slovenian Japanese students. While further research is indeed required in this area, the goals set at the beginning of the analysis have been achieved.

References

Corder, S. P. (1967). The Significance of Learner's Errors. *IRAL* 1967 (5), pp. 161-170.

Harasawa, I. 原沢伊都夫. (2012). Nihongo sho chūkyū gakushū-sha no sakubun shidō: Gakushū-sha no goyō bunseki o moto ni [日本語初中級学習者の作文指導：学習者の誤用分析をもとに] (Composition learning for learners of Japanese on the basic and intermediate level, based on an analysis of learner errors). *Shizuokadaigaku kokusai kōryū sentā kiyō* 静岡大学国際交流センター紀要, 6, pp. 79-92. Accessed 2. 9. 2018. https://ci.nii.ac.jp/naid/110008917835

Ichikawa, Y. 市川保子. (1993). Chūkyūreberu gakushūsha no goyō to sono bunseki - fukubun kōzō shūtoku katei o chūshin ni [中級レベル学習者の誤用とその分析─複文構造習得過程を中心に─] (The errors of students on the intermediate level – with focus on the process of acquisitions of compound sentence structures). *Nihongo kyōiku* 日本語教育, 81, pp. 55-66.

Ichikawa, Y. 市川保子. (1997). *Nihongo goyō reibun shōjiten [日本語誤用例文小辞典] (Small dictionary of examples of misuse in Japanese)*. Tokyo: Bonjinsha.

Kawaguchi, R. 川口良. (1995). Chūjōkyū nihongo gakushūsha no sakubun ni miru goyō no ichirei [中上級日本語学習者の作文にみる誤用の一例] (Types of errors that appear in the compositions of learners of Japanese on the intermediate and advanced level). *Gengo bunka to nihongokyōiku 言語文化と日本語教育*, pp. 178-188.

National Institute for Japanese Language and Linguistics. (2016). Learner Corpus Study of Aquisiton of Japanese as a Second Language. *NINJAL*, http://lsaj.ninjal.ac.jp/, Accessed 10. 4. 2018.

Noda, H. 野田尚史, & Sakoda, K. 迫田久美子. (2019). *Gakushūsha kōpasu to nihongo kyōiku kenkyū 学習者コーパスと日本語教育研究 (Learners' Corpora and Japanese Language Education Research)*. Tokyo: Kurosio.

Otsuka, K. 大塚薫, & Masayoshi, H. 林翠芳. (2010). Chū jōkyū reberu no Nihon gogakushūsha no sakubun shidō — iken bun ni miru goi kanji shiyō oyobi goyō no bunseki kekka o fumaete — [中上級レベルの日本語学習者の作文指導—意見文にみる語彙・漢字使用及び誤用の分析結果を踏まえて—] (Teaching composition of Japanese language learners at middle and upper level-based on analysis of vocabulary, kanji use and misuse in opinion sentences). *Kōchidaigaku sōgō kyōiku sentā shūgaku ryūgakusei shien bumon kiyō 高知大学総合教育センター修学・留学生支援部門紀要,* 4, pp. 47-66. Accessed 2. 9. 2018. https://ci.nii.ac.jp/naid/120002187909

Suzuki, T. 鈴木智美. (2002). 2000-nendo chūkyū sakubun ni mirareru goi imi ni kakawaru goyō — sho chūkyū reberu ni okeru goi imi kyōiku no jūjitsu o mezashite [2000 年度中級作文に見られる語彙・意味に関わる誤用—初中級レベルにおける語彙・意味教育の充実を目指して—] (Misuse of vocabulary and semantics found in the composition of students on the intermediate level in the year 2000 - Aiming at enhancement of vocabulary and semantics education on the beginner and intermediate level -). *Tōkyōgaikokugodaigaku ryūgakusei nihongo kyōiku sentā ronshū 東京外国語大学留学生日本語教育センター論集*, 28, pp. 27-42. Accessed 2. 9. 2018. http://repository.tufs.ac.jp/bitstream/10108/20943/1/jlc028003.pdf

Teramura, H. 寺村秀夫. (1990). *Gaikokujingakushūsha no nihongo goyōreishū* [外国人日本語学習者の日本語誤用例集] (Collection of misuse of foreign Japanese learners). Teramuragoyōreishū database 寺村誤用例集データベース. Accessed 15. 1. 2018. http://teramuradb.ninjal.ac.jp/teramura.goyoureishu.pdf

Umino, T. et al. (2012). Learners' Language Corpus of Japanese. *Tokyo University of Foreign Studies*. Accessed 1. 9. 2018. http://cblle.tufs.ac.jp/llc/ja/index.php?menulang=en

Yō, H. 楊帆. (2014). Chūkyū Nihongo gakushūsha no sakubun ni okeru konnan-ten: Bun kōzō no koōkankei ni tsuite [中級日本語学習者の作文における困難点：文構造の呼応関係について] (Difficulties in the compositions of Japanese learners on the intermediate level: on correspondence of sentence structure). *Akitadaigaku kokusai kōryū sentā kiyō 秋田大学国際交流センター紀要*, 3, pp. 15-28. Accessed 2. 9. 2018. https://ci.nii.ac.jp/naid/110009768148/en/