NAIST-IS-DD0661003

Doctoral Dissertation

Automatic Error Tag Annotation on the Writing of Japanese Language Learners for Linguistic and Educational Research

Hiromi Oyama

March 14, 2016

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Hiromi Oyama

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Kenichi Matsumoto	(Co-supervisor)
Associate Professor Masashi Shimbo	(Co-supervisor)
Associate Professor Mamoru Komachi	(Co-supervisor)
Assistant Professor Hiroyuki Shindo	(Co-supervisor)

Automatic Error Tag Annotation on the Writing of Japanese Language Learners for Linguistic and Educational Research*

Hiromi Oyama

Abstract

Recently, various types of learner corpora have been compiled and utilized for linguistic and educational research. As web-based application programs have been developed for language learners, a large size of language learners' texts is able to be collected on the web. These learner corpora include not only correct sentences but also incorrect sentences. Our object is to take advantage of these incorrect sentences for linguistic and educational research. In language education field, the researchers and language teachers wish to investigate the mechanism why learners make such errors, for leaners not to make the same mistake again and to use the insights learned from such corpora. However, it is not an easy task to process large corpora without any annotation nor any software to search in them. In order to make use of the corpora for those research, it is required to extract the errors in them, to add useful information and to learn from the insights appearing in the real use.

To this end, this study aims to do several tasks regarding learner corpora facilitation. The tasks are listed below.

1. To construct an error-tagged corpus (the NAIST Goyo corpus) for educational research.

   (a) To construct reliable error types for language learners and teachers.

*Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0661003, March 14, 2016.

(b) To use the corpus to investigate a particle usage of Japanese language learners. The result shows that particle omission is the most frequent error type and especially “no” and “wa” are the most difficult among all particles.

2. To investigate an approach to classifying incorrect sentences according to their error types. There is no such work done in the texts of learners of Japanese (hereafter LJ) so far and the experiment results 80 points in precision, which leads to realize an automatic error tag annotation application.

   (a) To apply an error type classification task to an out-of-domain text since there is no inter-corpus evaluation on error type classification task.

   (b) The experiment on out-of-domain corpus shows a lower accuracy than the in-domain text by 14.9 points.

3. To create an classification model for the usage of “wo” with newspaper corpus. The appropriate model of “wo” is applied to a learner corpus in order to distinguish an error sentence from an correct sentence. In the 100-instance test set, the result shows 50 points in F scores and in the 200-instance test set, it shows 53.9 points.

We have found that a new methodology for language education research through the learner corpora development.

Keywords:

Learner Corpora, Error Annotation, Error type classification, Learners of Japanese, Automatic Error Detection

日本語学習者の作文コーパスの言語教育研究のための誤用タグアノテーションの自動化*

大山 浩美

内容梗概

近年，様々な種類の言語学習者コーパスが収集され，言語教育の調査研究に利用されている．ウェブを利用した言語学習アプリケーションも登場し，膨大な量のコーパスを収集することも可能になってきている．学習者が生み出した文には正用だけでなく誤用も含まれており，それらの大規模な誤用文を言語学や教育などの研究に活かすことが重要である．日本語教育の現場では，学習者の書いた作文において学習者が誤りを犯す原因を追及し，誤用を犯さないようにフィードバックとして活かしたい需要があるが，大規模な言語学習者コーパスを調査分析するのは困難である．研究に活かすために，学習者コーパス内の学習者の書いた誤用を検出し，それらに誤用の種類を明示する誤用タグを振り，統計的に分析することが重要である．そのような理由から，本研究では以下の作業，実験を行い，以下のようなことが分かった．

1. 日本語学習者の作文に誤用タグを付与した誤用コーパス (NAIST 誤用コーパス) を作成した．

   (a) 現存する誤用タグの長所短所を考慮し，汎用性があるような誤用タグを作成した．

   (b) NAIST 誤用コーパスを用い，日本語学習者の格助詞の誤用について調査した．その助詞誤用頻度分析の結果，助詞を脱落させる誤りが最も多いことがわかった．さらに，助詞「の」や「は」において学習者の習得の難しさが見られた．

*奈良先端科学技術大学院大学 情報科学研究科 博士論文, NAIST-IS-DD0661003, 2016 年 3 月 14 日.

2. NAIST 誤用コーパスにおいて機械学習法を用いた誤用タイプ別自動分類実験を行い，かつアプリケーションに堪えうる適合率を実現した（約 8 割）.

   (a) 誤用タイプ別自動分類実験をドメイン外のコーパスにおいても行った.

   (b) NAIST 誤用コーパスでの実験に比べると精度が 14.9 ポイントほど低かった.

3. 新聞コーパスにおいて格助詞「を」の正用例，誤用例抽出実験を行った.

   (a) 新聞コーパスにおいて，格助詞「を，に，が，で，と」の正用モデルの抽出実験を行った.

   (b) 各助詞「を」の正用モデルを学習者コーパスで誤用例をはんていできるかどうかの実験を行った．100 事例の場合，F 値で 50.0 ポイント，200 事例の場合，F 値で 53.9 ポイントの精度で判定できた.

以上の作業を行い，学習者コーパスの整備をすることにより，言語教育の調査研究に関して新しい知見が得られた.

キーワード

学習者コーパス，誤用アノテーション，誤用タイプ自動分類, 日本語学習者, 自動誤用検出

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advance of data storage and computer processing technology, the linguistic resources for the research have been growing. Language teachers and researchers need a huge body of texts such as a newspaper corpus, a web-text corpus and a corpus of language learners writing as linguistic resource. Since learner corpora consist of language learners' spoken or written texts and are a valuable resource for reconsidering teaching methodology, materials or classroom management, they have been receiving attention for linguistic and educational use.

Dagneaux [8] use corpora for error analysis on the writings of learners of English as a Foreign Language (EFL). Granger [16] analyzes the uses of tenses by advanced learners of EFL with an error-tagged corpus. Such corpora offer researchers findings based on the fact from a different angle. Learner corpora can also provide positive and negative examples that contribute to improved writing skills, offer teachers with effective feedback on patterns of errors repeatedly made by students [12]. To master a foreign language, it is very effective to see why learners make a mistake and what causes it, rather than merely learning the correct expressions. It helps learners to store the contents they have learnt into their memorization system with concrete examples.

However, learner corpora include not only correct sentences but also incorrect sentences. Those incorrect sentences consist of different types of errors, which can be grammatical, semantic, stylistic, spelling errors and so forth. In order to use learner corpora properly, several pre-processings are needed to put those incorrect sentences in use for linguistic and educational research.

As for the natural language processing field, automatic error detection has been

actively studied. Since error types are too numerous to detect, some researchers have broken down the error detection task according to the types of errors in the texts, such as ill-formed spelling errors [31, 50], mass count noun errors [3, 35], preposition errors [7, 10, 11, 15, 48] and article errors [11, 15, 18]. Instead of addressing specific error types, Sun et.al [43] focus on discriminating between incorrect and correct sentences without considering error types.

As for texts by Japanese language learners, most of research focus on the particle (or postposition) error correction [21, 22, 36, 38, 41, 44]. Besides, Mizumoto et al. [33] perform error correction with a machine translation method for all error types in learners' writing.

Swanson and Yamangil [45] study error type classification over learner corpora of English and deal with 15 error types in the essays in the Cambridge Learner Corpus (CLC[1]). However, they did not report an inter-corpus evaluation.

In this thesis, the main contributions are as follows:

1. To construct an error-tagged corpus (the NAIST Goyo corpus) for educational research.
   (a) To construct reliable error type categorization.
   (b) To use the corpus to investigate a particle usage of Japanese language learners. The result shows that a particle omission type is the most frequent error and especially "no" and "wa" are the most difficult to learn among all particles.
2. To investigate an approach to classifying incorrect sentences according to their error types. There is no such work done in the texts of learners of Japanese so far and the experiment results in 80 points precision, which leads to realize an automatic error tag annotation application.
   (a) To apply an error type classification task to an out-of-domain text since there is no inter-corpus evaluation on error type classification task.
   (b) An experiment on out-of-domain corpus shows a lower accuracy than the in-domain text by 14.9 points.

[1] http://www.cambridge.org/elt/corpus/clc.htm

3. To create a classification model for the usage of “wo” classification with a newspaper corpus.

   (a) An unsupervised model of “wo” is applied to a learner corpus in order to distinguish an error sentence from a correct sentence. In the 100-instance test set, the result is 50 points in F scores and in the 200-instance test set, it achieves 53.9 points.

This thesis describes constructing a Japanese learner corpus, also illustrates the error types and the statistics on particle usage in the NAIST Goyo corpusin Chapter 2 . Chapter 3 describes error type classification on the NAIST Goyo corpus. Chapter 4 describes Japanese particle error detection task and includes previous research on automatic error detection and classification tasks. Chapter 5 concludes this work and explains future directions.

Chapter 2

Error Type Classification for Japanese Learners' Corpus

2.1. Introduction

This section discusses how to define error types and construct an error tag set corresponding to the error types. There are several Japanese learners corpora with error annotation. Since the size of each of these corpora is rather small and the error type schema of each corpus is also different, it is difficult to use all of them together. In addition, their annotation purposes are not effectivey applicable for machine learning but for linguistic research. The agreement rate between annotators is also not reported to know how reliable the annotations on the corpora are. Our work solves these issues residing in the corpora.

First, we investigate the characteristics of English learner corpora and Japanese language corpora and their error tag sets. Second, the data and the methods for the error type classification and the outline of the experiment is explained. Next, the features for the classification is illustrated. Lastly, we examine the result and analyze the possible causes of unsuccessfully classified instances.

2.2. Previous Work

Japanese Learner Corpora:

There are several Japanese learner corpora such as Taiyaku DB, which is a multilingual database of Japanese learners' essays compiled by the National Institute of Japanese Language (NINJAL)[1]. It consists of 1,565 essays written by learners from 15 different countries[2]. KY corpus [26] has spoken data of Japanese language learners at different proficiency levels. The corpora have different error type scheme which seem difficult to apply for machine learning and the agreement rate between annotators are not checked.

There are several Japanese language learners' corpora with error annotation, such as the Teramura corpus at Osaka University [46] (3,131 sentences with error tag annotations among 4,601 sentences), the learner corpus at Nagoya University [39] (756 files), the Online Japanese Error corpus dictionary[3] (whose files are error-tagged) and the Japanese learners' written composition corpus at Tsukuba University [27][4] (540 files). The Tsukuba corpus has only three kinds of error tags such as grammar, spelling and styles. Corpus of Chinese Learners of Japanese at Dalian Polytechnic University (henceforth Dailan Chinese corpus) is also error-annotated. Since the size of each of these corpora is rather small and the error schema of each corpus is also different, it is difficult to use all of them together. Since their annotation purposes are for linguistic research, it appears not very applicable for machine learning. To be specific, the agreement rate between annotators is not reported to know how reliable the annotations on the corpora are. Thus, the error-annotated corpus, the NAIST Goyo corpus, is created for our experiment, which is explained in section 2.4. The Taiyaku DB, Nagoya learner corpus, Dailan Chinese corpus are illustrated in detail in section 2.3.

English Learner Corpora:

ICLE (International Corpus of Learner English) corpus[5] is collected at University

[1] http://jpforlife.jp/contents_db

[2] There are several versions in Taiyaku DB according to the year of compilation. We used the first version of Taiyaku DB that consists of essays written by learners of 8 countries. The first version was only available when the experiment was conducted.

[3] http://cblle.tufs.ac.jp/llc/ja_wrong/index.php?m=default

[4] http://www34.atwiki.jp/jccorpus/

[5] http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cecl-Projects/Icle/icle.htm

of Louvain-la Neuve and consists of English learners' writing from 14 different countries. JEFLL corpus (Japanese EFL Learner) is a English writing corpus consisting a free writing by 10,000 Japanese junior high and high school students. CLC (Cambridge Learner Corpus)[6] [37] has writing of learners of 75 different mother tongues and 20,000,000 words. Free Text Corpus [17] compiles writing of French leaners of English at the intermediate through advanced levels. JLE (Japanese Learner English) [23] by NICT (National Institute of Information and Communications Technology) consists of interview tests of 1,281 English learners of Japanese. MELD (Montclair Electronic Language Database) [13] collects and annotates text written by all levels of second language learners. The database contains 44,477 words of annotated texts. These corpora are illustrated in detail in section 2.3.

2.3. Existing Error Tag Set Construction

Each of the existing error tag set has unique aspects according to the purpose of each research. Thus, there are a number of error tag sets and it is difficult to choose one from them and also to merge all of them.

Error Tags in English Learner Corpora: There are three main characteristics in constructing error tag set: (1) to construct error tags with linguistic perspective [7]; (2) to see how the word has changed from the original errors; (3) to keep correction information. The first one indicates that the error should be analyzed linguistically such as morphologically, syntactically, phonologically and so forth. The second one indicates that the error is omitted, added unnecessarily or wrongly chosen. The third one indicates that the error tag should contain the correction. Table 2.1 shows a summary of the error tag sets used in English language learners' corpora. Each corpus is explained in detail below.

ICLE (International Corpus of Learner English) corpus has English learners' writing, each of which consists of 500 to 1,000 words and the entire corpus size is 200,000 words in total. Annotation is added onto 150,000 words among 200,000. Error tag in the example below is "GVT" and each capitalized alphabet shows error categories.

[6] `http://www.cambridge.org/elt/corpus/clc.htm`

[7] This number associates the number in Table 2.1

Table 2.1. Error tag set in English language learners corpora (Adv. indicates advanced learners and Bgn. indicates beginners.)

	ICLE	JEFLL	CLC	Free Text	NICT JLE	MELD
Year	1998	2000	2003	2003	2004	2005
Author	Dagneaux	Tono	Nicholls	Granger	Izumi et al.	Fitzpatrick & Seegmiller
1	✓	✓	✓	✓	✓	-
2	✓	-	✓	✓	-	-
3	✓	✓	✓	✓	✓	✓
Target	English	English	English	English	English	English
Levels of students	Adv.	Bgn.–Adv.	Bgn.–Adv.	Bgn.–Adv.	Bgn.–Adv.	Adv.
Size of corpus	2M	700K	2M	450K	2M	100K

*The numbers in 3rd to 5th line associate the characteristics of error tag set.

"G" indicates "Grammar error" in the main category, "V" indicates "Verb" in Part-Of-Speech (henceforth POS) category and "T" in linguistic aspect as tense. The words enclosed with "$" are correction to the previous error words. The learner is supposed to write "Barons that had lived in those castles." instead of "Barons that lived in those castel." Tag is not a closed style [9].

ICLE :

- `Barons that (GVT)lived$had lived$ in those (FS)castel$castles$.`

JEFLL corpus (Japanese EFL Learner) describes both correct and error sentences. "ER_ART" is a tag of article error and "ART" is its correction to the error. The tag is a closed style.

JEFLL :

- (correction) `I have hardly had <ART>a</ART> bad dream.`

- (error) From <ER_ART>the</ER_ART> cliff.

CLC (Cambridge Learner Corpus) has an error annotation on 5,000,000 words. In the first example below, “U” in the triangle brackets indicates “Unnecessary” and “A” indicates “Pronoun”, which mentions “they” in this sentence is not necessary. The second example says that “You hardly ever meet people ...” is incorrect and the learner is supposed to write “Hardly ever do you meet people ...”. Error tag indicates an error of argument structure (AS) and a closed style.

CLC :

1. Lawyers, doctors, etc,<#UA>they</#UA> hardly earn $50,000 a year.
2. <#AS>Hardly ever do you meet:You hardly ever meet</#AS>people...

Free Text Corpus [17] is error-annotated on 300,000 words out of the entire 450,000 words. Tags are based on the linguistic analysis and a POS information and include the correction to the error. In the example above, “F” indicates that a formal error which includes a notation error, capitalization error or spelling mistakes. “DIA” indicates a diacritic error which is a particular spelling error often seen in French texts. “NOM” shows a POS information about “noun”. The example says “secret” is a diacritic error and its correction is “secrét”.

Free Text :

- ...qui ne sait pas garder le moindre <F><DIA><NOM>#secret$secrét </NOM></DIA></F>.

NICT JLE (Japanese Learner English) consists of transcribed interview tests, each of which lasts 15 minutes per person. The entire corpus size is 2,000,000 words, among which 167 people’s files are error annotated. The tags contain POS information, grammatical mistakes and lexical mistakes. The error taxonomy of this corpus has POS information in its first stage such as noun, verb, auxiliary, adjective, adverb, preposition, article, pronoun, conjunction, relatives and interrogatives. Under each of these POS category is there conjunction error, word choice error and so forth. Thus, each

POS information includes more specific error types of above, which have a possibility to confuse the annotators.

Since the POS category in JLE corpus is an important factor for the automatic error type classification task and the automatic error detection task that is mentioned later, these categories are used for our error types. In the example above, “n_num” inside the tag indicates that “n” is noun and “num” is a counting error either singular or plural noun. “crr=teams” inside the tag indicates that the correction to the error is “teams”.

NICT JLE :

- `I belong to two baseball <n_num crr="teams">team</n_num>.`

MELD (Montclair Electronic Language Database) [13] has no implicit tags but it uses the strings of an error and a correction as codes. With these tags, they can reduce the discrepancy and the misclassification between annotators. In the example above, “is” is an error and “are” is chosen as its correction. “0” indicates of an omission of word. The error tag set for this research is based especially on the JLE and MELD tag set.

MELD

- School systems {is/are} since children {0/are} usually inspired becoming {a/0} good citizens.

Error Tags in Japanese Learner Corpora:
Unlike English learner corpora, the designing of the error tag in Japanese learners corpus has two characteristics [42] .

1. To construct the error types according to linguistic description.
2. To construct the error types according to the actual errors seen in learners corpus.

English learner corpora mostly accept the latter method, however, Japanese learner corpora also have those both ways. The former method is effective when the first language of learners and types of the errors are various [42]. The error tag set of the

NAIST Goyo corpus is based on the NICT JLE corpus that is mentioned above, the learner corpus at Nagoya University [39] (756 files), the corpus of writing of Chinese learners of Japanese at Dalian polytechnic university [42], and "A Dictionary of Japanese Language Learners' Errors I and II [19, 20]".

Nagoya Learner Corpus:

Nagoya learner corpus [39] is constructed with the linguistic description (1) and based on Masuoka & Takubo grammar [29]. However, they have no classification such as "Addition, Omission, Replacement or Position" unlike [19, 20].

The example[8] below is taken from Nagoya leaner corpus. These tags are extended from JCHAT tag created to annotate children's utterance. The error strings are surrounded with triangular brackets and the indexed [∗] is also added. "%err" is called "correction tia" and includes the error and its correction. "%als" includes the error analysis. "%err" and "%als" are indexed in order to align "%err" and "%als" with the target error strings. There are also "%com" in which annotators comments are added.

- ∗GAK:<一般的に添加物は自然な物じゃなくて，また子供達は小さいから>
  - ∗GAK:<ippantekini tenkabutsu wa shizennamono janakute, mata kodomotachi wa chiisaikara>
  - ∗GAK:<Since the additives are not natural and the children are still small>
- [∗1] >[∗2] , <添加物を入っている >[∗3] 物をなるべく避ける.
  - [∗1]>[∗2] <tenkabutsu wo haitteiru>[∗3] mono wo narubeku sakeru.
  - [∗1]>[∗2]<the additive are added>[∗3](you) should avoid taking things
- %err: [1 また子供達は小さいから = まだ子供達は小さいから];
  - %err: [1 mata kodomotachi wa chiisai kara = mada kodomo tachi wa chiisai kara];

[8]This example is taken from http://cookie.nagoya-u.ac.jp/pub/goyooman.html, however, it is no longer able to access. It is moved to http://lang.nagoya-u.ac.jp/~sugiura/CHILDES/goyooCHILDESformat.html with a slight change in explanation.

- %err: [1 the children are still small];

▹ [2 一般的に添加物は自然な物じゃなくて，まだ子供達は小さいから = 一般的に添加物は自然な物じゃないし，まだ子供達は小さいから];

- [2 ippantekini tenkabutsu wa shizenna mono janakute, mada kodomo tachi wa chiisai kara = ippantekini tenkabutsu wa shizenna mono janaishi, mada kodomotachi wa chiisaikara];
- [2 Since the additives are not generally natural and the children are still small.];

▹ [3 添加物を入っている = 添加物が入っている];

- [3 tenkabutsu wo haitteiru = tenkabutsu ga haitteiru];
- [3 the additives are added];

▹ %als: [1 濁音が清音の表記になっている。「た」];

- %als:[1 dakuon ga seion no hyouki ni natteiru. “ta”];
- %als:[1 “ta” should be spelled with the voiced consonant “da” instead of the unvoiced consonant.];

▹ [2 テ形による接続の間違い。接続助詞「し」を使ったほうがいい。];

- [2 tekei niyoru setsuzoku no machigai. setsuzokujoshi “shi” wo tukatta houga ii.];
- [2 Conjunction error using of te form. Conjunction particle “shi” should be used.];

▹ [3 格助詞「を」と「が」の間違い。「入っている」の主体];

- [3 kakujoshi “wo” to “ga” no machigai. “haitteiru” no shutai.];
- [3 Confusion in particle “wo” and “ga”. “tenka butsu (additive)” is the subject of the verb “haitteiru”];

Corpus of Chinese Learners of Japanese at Dalian Polytechnic University (henceforth Dailan Chinese corpus):
The Dailan Chinese Corpus adopts the error types based on the knowledge from the actual errors the learners made [42]. They created their own error tags based on the errors seen in the corpus. Since their error taxonomy is based on the actual errors, the more unique aspects are seen such as the errors of demonstrative, formal noun, counters and Chinese origin words than [20, 39, 19]. Their tag set describes target modification (or formality classification) and linguistic characteristics. Target modification indicates "Addition", "Omission", "Confusion", "Misordering", "Misformation" and "Transfer". Linguistic characteristics describe the linguistic analysis such as tense, aspect, extra postpositions, confusion between "wa" and "ga" and etc. The example below is taken from the Dailan Chinese corpus.

- ▹ 今日本語の専攻し (→ 日本語を専攻し) ながら，コンピュータも独学している．
- Ima nihongo no senkou shi (→ nihongo wo senkou shi) nagara, konpyuutaa mo dokugakushiteiru.
- I am studying computer while taking a Japanese course.

Table 2.2. The Error Taxonomy in [19]

Main category	Sub category
Mood	20
Tense/Aspect	10
Intransitive/Transitive verb/Voice	5
Giving/Receiving	3
Postposition “wa”	3
Particles/Attributive particle/Compound particle	10
Continuous/Adnominal modification	2
Subordinate clause	2
Total	86

Ichikawa’s Japanese Error Corpus:

The error types in [19, 20] are “Mood” (also known as “Modality”), “Tense/Aspect”, “Intransitive/Transitive verb/Voice”, “Giving/Receiving”, “Postposition wa”, “Particles/Attributive/Compound particle”, “Continuous/Adnominal modification”, “Subordinate clause” in Table 2.2. However, “Replacement, Omission and Addition” in [19, 20] are located under every main categories.

- ▹ Particle, Addition
- ▹ 兄弟は８人が (→ϕ) いて，シアトルやシカゴに住んでいる.
- Kyoudai wa 8 nin ga (→ϕ) ite shiatoru ya shikago ni sundeiru.
- I have 8 siblings who live in Seattle and Chicago.

In the example above, “ga (→ϕ)” indicates an omission error of “ga” particle. Ichikawa [19, 20] analyzed learners errors, however, she did not annotate them with error tags.

The NAIST Goyo corpus consists of 76 error types based on those existing error types [23, 19, 20, 42]. Those error type description in detail are added in Appendix B.

Shimizu et. al [42] possess some unique tags we mentioned above. We included these tags since we found these tags appropriate for our error types. However, the errors of Japanese postposition “wa/ga” in Dailan Chinese corpus are separated from

Table 2.3. Error Tag Set on Japanese Language Learners Corpora

	Oso et al.	Ichikawa	Shimizu et al.
Year	1997	1997	2004
Linguistic description	✓	✓	✓
Target modification	-	✓	✓
Correction	✓	✓	✓
Designing method	Linguistic theory	Data driven	Data driven

particle errors even though "wa/ga" are considered as one of postpositions. They have chosen which error types should be selected with their own research interests. Ichikawa [19, 20] also has an individual category for postposition "wa". However, we integrated the "wa/ga" error type into "Particle" error type. In addition, since the Dailan Chinese corpus consists of writing of Chinese language learners, the error types particular to Chinese learners are not taken.

2.4. Error Tag Annotation on the NAIST Goyo corpus

2.4.1 Annotation Schema

Granger [17] summarized important points in constructing an error tag set. These points include consistency, informativity, flexibility and reusability in tags. From the points of the consistency and reusability, the target modification is most effective because there would be less disagreement between annotators than judging with the linguistic description. However, this target modification can provide less linguistic information for research than adding linguistic description. Whether words are omitted, added or replaced, is not as useful as the linguistic description for learners to learn why they make an error in writing. From the teachers' points of view, it is effective to use more linguistic reasoning than omission or addition. James [24] recommends an error tag set consisting both the linguistic description and the target modification. In addition, flexibility is also an important key because the tag should be used with any research perspectives. Thus, we consider that: the tag set should be consistent so that annotators can judge at the same standard from the beginning to the end; understandable so that they can agree with their judgement; informative so that teachers and

learners can use them for feedback; flexible and reusable so that any researcher can use the tag set for their own research purpose.

2.4.2 The NAIST Goyo Corpus

The NAIST Goyo corpus consists of error-annotated 313 essays from Taiyaku DB. They are annotated with error tags and other information, which already been corrected by professional teachers of Japanese. One benefit to use essays from Taiyaku DB is that they are written by a variety of the nationalities so as to take examples from a wider range of errors. Considering these characteristics of each error tag set, we constructed our own error tag set and annotated with the error tags and other information tags onto 313 files of writing from Taiyaku DB.

The NAIST Goyo corpus contains omitted strings such as "Nobu {*[9] ϕ **/ to iu**} restoran ni ikimashita (I went to the restaurant which is called Nobu.)". In this sentence, the error string "to iu" is difficult to be categorized to certain error type because "to iu" is composed of particle "to" and verb "iu". The string is sometimes too long and compositional to categorize. These sentences are categorized as "Omission" (or "Addition") in our error scheme. The NAIST Goyo corpus possess "Omission" and "Addition" under the main category "Postposition", even though they are a very few as mentioned under the "Omission" section in Appendix A. We constructed "Omission" and "Addition" as individual group, not as the sub-category.

Figure 2.1 shows an example of annotated texts from the Taiyaku DB about "smoking" by a Chinese learner of Japanese. A file is surrounded by <corpus> and the writer information is described on the top. Under the information tag, there are 6 kinds of individual information of <id> (index number), <name> (the learner's name), <gender> (gender), <nationality> (nationality), <m-lang> (mother tongue) and <year> (the year in which the writing is made). The entire writing is surrounded by <text> tag. Inside the text, paragraph is surrounded by <p>, sentence is surrounded by <s> and each error string is surrounded by <goyo>. Inside the error tag includes "type" attribute which indicates error type and "crr" attribute which indicates a correction to the error. All of the corrections in the NAIST Goyo corpus are the same as ones already added in Taiyaku DB.

[9]The asterisk mark indicates an error instance.

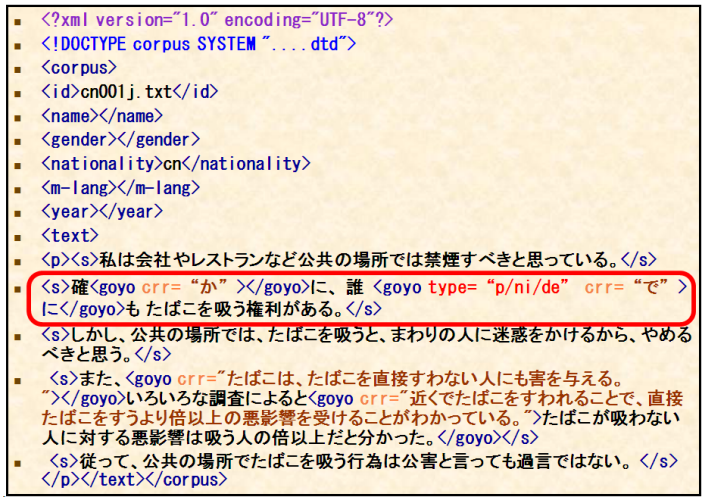


Figure 2.1. Example of Error-Annotated Sentence

<s>... 誰 <goyo type="p/ni/de" crr="で">に</goyo>もたばこを吸う権利がある。</s>

- Dare demo tabako wo suu kenri ga aru.
- Everybody has a right to smoke.

One example below is taken from the sentence in Figure 2.1 and “ni” is the target error string surrounded by <goyo>. The correction is “de” written as "crr" attribute inside the error tag and its error type is written as “p/ni/de” as "type" attribute[10]. “p” in the first category of the error type tag indicates main error type is a particle error. The error string is written in the second category and the correction comes in the third category. One of characteristics of this tag set is to include attributes such as "type" and "crr" (correction) inside the tag, which is similar to JLE corpus. Another characteristic is that the tag reflects how the error is corrected so that one can easily understand what the tag indicates by looking at the tag itself.

The other example is listed below. In the case of multiple errors in a sentence, it is tagged with "typeN"[11]. There is also a case that multiple ways of corrections are pos-

[10]These corrections are based on ones provided in Taiyaku DB.
[11]N is a counting number and its order is not considered.

Table 2.4. Linguistic Reasoning for Particle Tag

Error&Correction	tag	e.g.	reason
“wa” → ϕ	“p/wa/ad”	watashi wa nihon ni kita bakari no toki no aruhi wa → aruhi	There is no need of “wa” after the date
“wa”→“ga”	“p/wa/ga”	watashi wa nihon ni yattekita mokuteki wa → watashi ga nihon ni	Confusion between “wa” and “de”
“ni”→“de”	“p/ni/de”	nihonshakai ni seikatsu suru nowa mondai ga nai →n ihonshakai de	“de”: place for action
“ni”→“de”	“p/ni/de”	waapuro nino sousa wo minitsuketai → waapuro deno	“de”: for tools

sible depending on each errors, but we did not consider that matter in this annotation work.

<s>それで,<goyo type1="sem" type2="not/kj" crr="常に">まいにち</goyo>がいこくのえんじょがいります.</s>

- Sorede, tsuneni gaikoku no enjo ga irimasu.
- And, (we) need an international help all of the time.

In the example above, the error string, “mainichi (everyday)”, is tagged with <goyo> and the corrected string, “tsune ni (all of the time)”, is with "crr". Two ways of corrections should be made onto this sentence such as “mainichi (everyday)” to “tsune ni (all of the time)” and then “tsune ni (all of the time)” to the kanji version. Two error types are added such as “Word choice (SEM) (type1="sem")” for the first correction and “Spelling (NOT) (type2="not/kj")” for the second correction.

Agreement between Annotators

The agreement rate are investigated between two annotators who added error tags onto the NAIST Goyo corpus. They have been working as annotators more than 5 years. They are asked to annotate the same text with the error tags in the NAIST Goyo corpus. κ score [4] is calculated in order to know how much they assign the same tags to the same errors. 170 sentences are randomly selected; 10 sentences from each error type, which is 1.2 points of the total number of instances. They are also asked to choose the first choice and the second choice of the error type. The agreement rate shows $\kappa = 0.602$ at the first choice and $\kappa = 0.654$ at the second choice. When the κ score ranges from 0.81 to 1.00, it indicates "almost perfect" match. When the κ score ranges from 0.61 to 0.80, it indicates "practically" match. Thus, we consider the κ shows "practically" match of the annotation on the NAIST Goyo corpus. This score contributes the reliability of the annotation in the error-annotated corpus. In order to pursue the reliability, the disagreement between annotators should be minimized as less as possible. If the agreement rate is high, the reliability of the error tags is increased. In the NAIST Goyo corpus, the annotators can annotate not multi-class tags but multi-label tags without knowing the hierarchical construct of the error types.

2.4.3 Postposition Error in the NAIST Goyo corpus

The postposition[12] errors in the NAIST Goyo corpus are tagged and analyzed statistically. The number of all postposition errors are 3,037 and that of type token is 791 (Table 2.5). Figure 2.2 represents 10 most frequent postposition errors in the NAIST Goyo corpus.

The number of omission error type is far more than other types. "ga" error type follows and then "wa", "wo", "no", "ni", "de", "mo", "toka" and "to" (Figure 2.2).

Among omission error type, "no" is the most frequent type (See the column of "p/om/no" in Figure 2.3 [13]). The frequent errors of "no" include of missing "no" after

[12]The Japanese language uses postpositions, joshi, to denote the direction of an action and who is performing the action. They consistently come after the word that they modify. Among the postpositions, there are particles which carry cases such as "ga", "wo", "no", "ni", "e", "de", "yori", "kara" and "to". "Wa" is called binding particles or adverbial particles, kakarijoshi. We differentiate postpositions and particles in this paper.

[13]"sim" in 6th column of Figure 2.3 indicates punctuation marks.

Table 2.5. Number of Postposition Error in Taiyaku DB

	token	type
China (cn)	179	64
Cambodia (kh)	92	53
Korea (kr)	722	171
Malaysia (ml)	179	182
Mongolia (mn)	80	39
Singapore (sg)	67	35
Thai (th)	579	112
Vietnum (vn)	413	135
Total	3,037	791

counters and between two-kanji compound words. This is an example of omitted “no” after counters by a Chinese learner of Japanese.

▹ cn069j.txt

- <s>二匹<goyo type="p/om/no" crr="の"></goyo>黄牛と一緒にお互い<goyo type="p/ni/wo" crr="を">に</goyo>頼り合いました。</s>
- Nihiki no kougyuu to issho ni otagai wo tayoriaimashita.
- (I) helped each other with two yellow cows.

The learner omitted “no” after “nihiki (two for counting animals).” Teacher corrected “nihiki kougyuu (two yellow cows)” into “nihiki no kougyuu (two yellow cows).”

The second most frequent type is an omission of “wa”. “Wa” is an adverbial postposition that is replaceable with any other postpositions. Thus, it is one of the most difficult postpositions to master for learners of Japanese.

▹ kr185j.txt

- <s>他人に<goyo type="col" crr="被害を与えている">がいする</goyo>と<goyo type="p/om/wa" crr="は"></goyo>おもいません。</s>
  - Tanin ni higai wo ataeteiru to wa omoimasen.

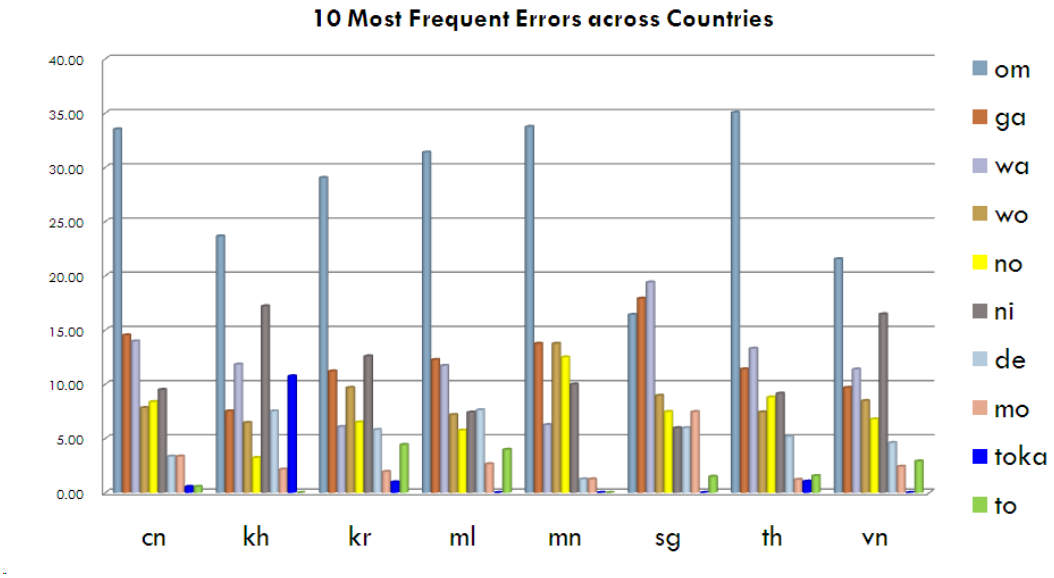


Figure 2.2. 10 Most Frequent Postposition Errors

– I don't think I do harm to others.

There are many noticeable errors of omission of "wa" in the sentences of "...to wa omoimasen (I don't think that ...)" or "...to wa iimasen (I won't say that ...)". "...to omoimasen (I don't think that ...)" even without "wa" is already grammatical and understandable, however, the teachers added "wa" in them. This use of "wa" is one of difficult usages of "wa" for learners to master and not fully researched that function. This error and correction pair is one of great insights in the actual errors found in the language learners' corpora.

2.5. Summary

The NAIST Goyo corpus is an error-annotated corpus created from Taiyaku DB compiled by NINJAL. Prior to adding the annotation, several existing learner corpora and

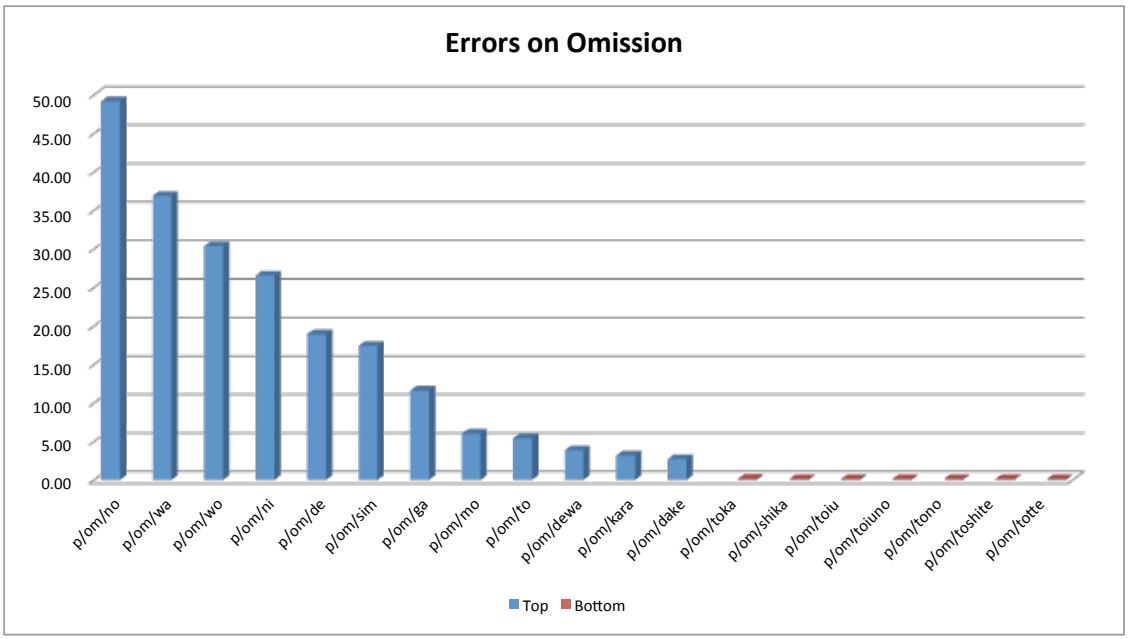


Figure 2.3. Frequency in Omission Error Type of Postposition

their error tag sets are investigated. Considering pros and cons, and characteristics in these previous work, we designed our own tag set and annotated onto Taiyaku DB. After those error tags are analyzed, it is found that omission type of postpositions is most frequent in the NAIST Goyo corpus. In addition, the usage of postposition errors of “no” and “wa” are the most difficult to master. Other error types such as “Verb” or “Word choice” are to be analyzed in the future work.

Chapter 3

Error Type Classification of Japanese Language Learners' Writing

3.1. Introduction

We aim at classifying sentences from learner corpora according to error types such as particle errors, verb errors and so forth. Error type classification benefits for linguistic research and education as well as the natural language processing technology. We made the machine learning-based approach to automatic error type classification on the writing of learners of Japanese (LJ). First, we investigate an approach to classify incorrect sentences according to their error types. Second, we also apply error type classification to an out-of-domain text. Finally, we discuss the instances correctly classified in this experiment.

3.2. Previous Work

As for automatic error type classification, Swanson and Yamangil [45] deal with 15 error types in the essays of learners of English in the Cambridge Learner Corpus（CLC [1]). However, they did not report an inter-corpus evaluation. Automatic error type classification in Japanese text have unexplored so far. Swanson and Yamangil [45] uses the existing English learners' corpora but they did not study about what error types

[1] http://www.cambridge.org/elt/corpus/clc.htm

suits for the error type classification task. They did not report about how the design of error types effects the classification task. In addition, they only use the string and parts-of-speech (POS) information from the error and correction. In our experiment, we use edit distance scores and substitution probability calculated from the web-based corpus of the Lang-8 corpus added onto those features.

3.3. Materials and Methods

3.3.1 Materials: Data

Error Types for the Experiment

To simplify this experiment, we utilized a compressed set of 17 essential error types out of 76 in total in the NAIST Goyo corpus. The 76 error types are summerized into 23 groups at the first category, from which we select the 17 essential error types. We also add some other useful error types for this experiment. By way of example, "Verb" takes in a "verb conjugation" error and the "Spelling" category includes Hiragana, Katakana or Kanji errors. Detailed is attached in Appendix A.

The 17 essential error types are illustrated as the top 17 error types in Table 3.1. The lower 6 error types in Table 3.1 are excluded for this experiment. Since "Modality" needs other types of features than we use here such as semantic features and subjectivity of a sentence. "Phrase" error includes the incorrect use of phrase patterns such as "...tari ...tari" in a sentence like "Kinou wa ne**tari** terebi wo mi**tari** shimashita (I took a nap and watched TV yesterday.).". "Phrase" type errors are excluded from this experiment. Since some of them are discontinuous such as "...tari ...tari", they are difficult to align the correct with incorrect sentences. "Whole alteration" indicates that the entire sentence needs rewriting. This type is also difficult to make an alignment. Thus, "Whole alteration" type errors are excluded from this experiment. Since "Miscellaneous" includes various error types and it is also not the scope of the experiment because each of them is too refined and the instance of these are small.

"Collocation", "Use of da", "Negation", "Adverb", "Pronouns" are included to our experiment even though the number of these instances are small. There are several research based on learner corpora such as Taiyaku DB or [26][2]concerning these error

[2]KY corpus is a spoken corpus and collects the interview texts from 90 leaners of Japanese of

types; “Use of da” ([25, 49]), “Negation” ([32, 34]), “Adverb” ([1, 2, 30]) and “Pronouns” ([5]). “Collocation” and “Negation” are also used for the error type classification[45]. “Collocation” needs a large body of text corpus in order to search the patterns that two or more words are seen at the same context [47]. We differentiate the word choice error and the collocation error by restricting the former type to the word unit errors. “Collocation” also includes not only “Nouns + particles + Verbs” but also “Adjectives + Nouns”. However, the second type of collocation are not our scope in this experiment.

Table 3.2 represents the proportion of error types according to the learners' national origin[3]. The most frequent error type is “Word choice (SEM)[4]”, followed by “Postposition (P)”, “Verb (V)”, “Spelling (NOT)”, “Phrase” and “Adjective (ADJ)”.

Human Judgement and Error Tags

Oyama [40] used the flat-structured error type according to the method that Swanson and Yamngil [45] used. However, we made a preliminary experiment so that the structure of the error type affects the classification performance.

We conducted an experiment with teachers of Japanese, having them classify errors to investigate how human judge the error types. We asked 11 Japanese teachers to classify 20 incorrect sentences randomly taken from the same test data according to the error types. After this experiment, we had an inquiring survey on these teachers. The findings from this experiment are listed below.

1. “SEM” and “V” are confusing error types.
2. Those Japanese teachers focus on the error strings, its correction and surrounding words.
3. They also focus on the dependency structure of those errors.

The most confusing error type was “SEM” followed by “V” as is shown in Table 3.3. The second row shows that “SEM” is mistaken as “P”, “V”, “STL”, “NOM”, “CONJ”, “ADJ”, “DEM”, “COL”, “AUX”, “NEG”, “ADV” and “PRON”. The teachers mentioned on the inquiring survey while they were classifying the sentences into the error types, when in doubt, they chose “SEM”.

English, Korean and Chinese speakers.

[3]The number is a proportion to the number of learners' essays.

[4]“Word choice” is abbreviated as “SEM” and so do other error types.

Table 3.1. The selected 17 error types (ϕ in this table indicates a missing element and # indicates the number of instances.)

Description	Sample and Correction	English Translation	#
Postposition (P)	*Eigo ***wo/ga** wakaru	I can understand English	3,351
Word choice (SEM)	{ ***bubun jin / ichibu no hito** }	**some people**	2,546
Spelling (NOT)	nen{***pa / pai**} no hito	the **elderly** people	1,838
Missing (OM)	Nobu {*** ϕ / toiu**} resutoran ni ikimashita	I went to a restaurant **called** Nobu	1,441
Verb (V)	tegami wo {***kaki / kaka**}nai	I **will not write** a letter.	1,348
Unnecessary (AD)	{ ***tenki ga /** ϕsamukute... }	**The weather** is so cold...	1,177
Inappropriate register (STL)	Totemo taihen {***ne / desu**}	It is very hard	328
Nominalization (NOM)	shumi wa eiga wo miru{***no / koto**} desu	I enjoy **watching** a movie	300
Connecting (CONJ)	{ ***Soshitemo / Soshite**} pet to asobimasu	**And then,** I played with my pet	196
Adjective (ADJ)	boku wa {***huto–kute / huto–i**}hito desukara	I am a **fat** person	149
Demonstrative (DEM)	{ ***Asoko / Soko**}de tomodachi ni aimashita.	I met a friend **there**	137
Word order (ORD)	{ ***yori shichigatsu / shichigatsu yori** }	**From** July	121
Collocation (COL)	Shiken {***ni sankashimashita / wo ukemashita**}	I **took a test**	113
Use of da (AUX)	Anohito wa kirei{***desu / da**}to omoimasu	I think **that** the girl is pretty	49
Negation (NEG)	Ie ni irare {***naide / nakute**} soto he ikimashita	I went out **because I did not** want to stay at home	26
Adverb (ADV)	Nonbiri {***ni / to**} sugoshita	I spend a day **at leisure**	24
Pronouns (PRON)	{ ***Karetachi / Karera** }	**they /them**	16
Nouns	*nichiyoubi wa gakkou ga {***yasumimasu / yasumidesu**}	**Sunday is a school holiday.**	10
Noun modification	{ ***Iwayuru /** ϕ} kokusaikankei	**so-called** international relationship	12
Modality	Sensei wa beiryugakushite ima wa tyugoku ni ikitai{***desukedo / soudesu**}	**I heard** my teacher wants to study in China.	59
Phrases	{ ***Otoko teare Onna teare**} { **Otoko deare Onna deare**}	**Regardless of** men or women	549
Whole alteration	{ ***fukai kankeishite tabako wo suu**} { **tabako wo suu noniwa hukai kankei ga aru**}	**There is a deep relation between smoking (and getting rid of stress).**	258
Miscellaneous	See Table 5.1		148

Table 3.2. The Proportion of Error Types in the NAIST Goyo corpus (top 10)
(VN indicates learners from Vietnam, TH Thai, CN China, ML Malaysia, MN Mongolia, KH Cambodia, KR Korea and SG Singapore)

	VN	TH	CN	ML	MN	KH	KR	SG
Word choice (SEM)	35.0	27.0	17.2	22.8	29.2	12.8	25.2	23.8
Postposition (P)	21.8	23.1	20.6	24.2	22.1	17.4	17.3	30.6
Verb (V)	13.8	15.3	16.8	12.1	14.2	15.9	14.6	10.2
Spelling (NOT)	9.8	10.1	19.8	16.9	12.7	33.6	15.5	6.8
Phrase	6.2	7.0	2.6	7.3	5.2	1.7	3.4	4.9
Nominalization (NOM)	2.5	2.6	3.5	1.4	3.4	2.0	4.4	2.9
Adjective (ADJ)	2.0	0.9	2.6	1.5	1.9	1.7	1.5	1.5
Whole alteration	2.0	2.6	1.2	3.4	0.7	1.4	2.4	2.4
Inappropriate register (STL)	1.7	1.2	2.3	6.0	4.1	6.1	3.1	6.3
Word order (ORD)	1.0	1.3	1.2	0.3	0.4	1.2	0.6	0.0

We considered the aspects the teachers focus on in classifying errors. We found that, though full sentences were provided, they judged error types mainly according to very local cues, such as, the error strings, its correction and sometimes surrounding words in a window size of 1. In addition, in the case of "P" errors, they tried to find the verb in dependency. In a similar way in case of "ADV", they tried to focus on the verb which the adverb depends on.

Figure 3.1 shows the tree-structured tag set containing all the essential tags of Table 3.1. In the first stage, all tags are divided into the three categories in the first place, "OM", "AD", Replacement. In the second stage, instances categorized as Replacement are classified either as Grammatical or Lexical. In the last stage, instances recognized as Grammatical and Lexical are classified by Grammatical and Lexical error classification models, respectively. Grammatical instances are grammar-based types where students err due to a lack of practice with the grammar. They are classified into the 12 categories of "P", "V", "STL", "NOM", "CONJ", "ADJ", "DEM", "ORD", "AUX", "NEG" , "PRON", "ADV". Lexical group contains dictionally-based types where students err due to a lack of lexical knowledge. They are classified as "SEM", "NOT" or "COL".

As a test data, 1,090 incorrect sentences are used from the Lang-8 corpus for an out-

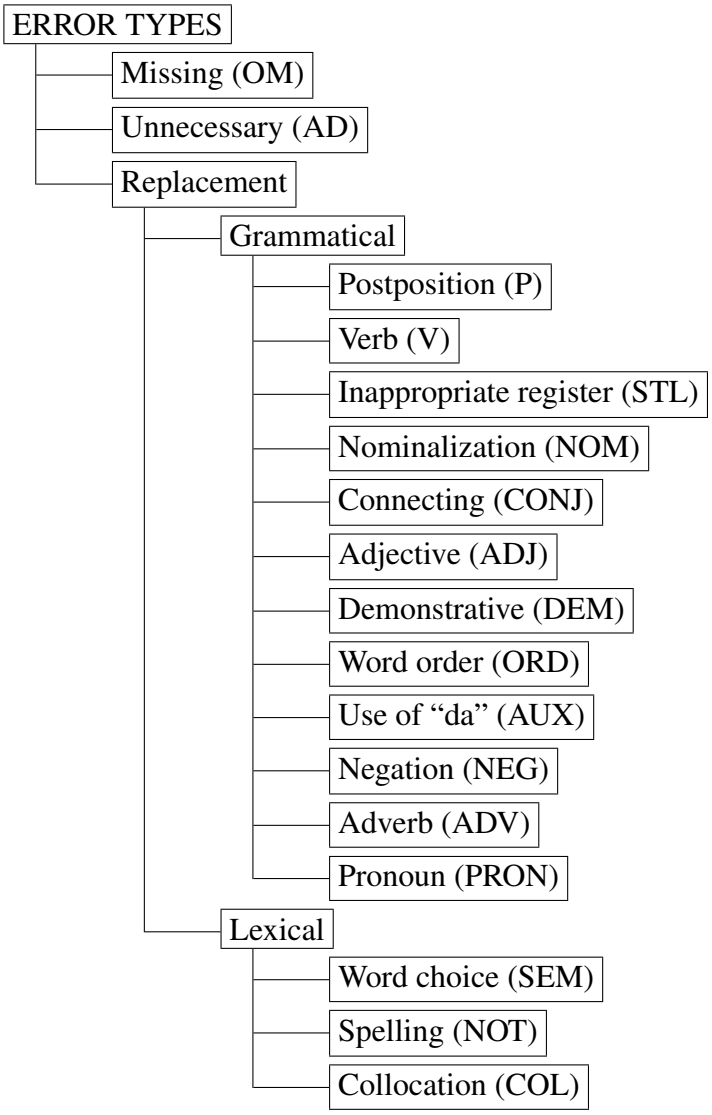


Figure 3.1. Tree-Structured Tag Set

Table 3.3. Confusion Matrix of the Human Judge over Error Type in Lang-8
(Row represents the actual classes and column represents the classes predicted by the teachers.)

	P	Sm	Nt	O	V	A	St	Nm	Cj	Aj	D	Or	Cl	Ax	Ng	Av	Pr
P	1	0	0	0	1	0	0	2	0	0	0	0	0	0	0	2	0
SEM	1	4	0	0	3	0	1	6	4	3	1	0	4	2	3	9	3
NOT	0	0	6	0	0	0	1	0	0	0	0	0	1	0	0	0	0
OM	0	0	2	0	0	1	0	1	1	1	0	0	0	3	0	0	0
V	0	2	0	1	7	0	0	3	5	0	0	0	1	5	3	0	0
AD	1	0	0	0	0	2	2	1	1	0	1	0	0	1	1	1	0
STL	2	0	0	0	0	0	5	1	0	0	0	1	0	0	0	0	0
NOM	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0
CONJ	0	0	0	0	0	0	1	0	5	0	0	0	0	3	1	1	0
ADJ	0	0	0	0	0	0	0	0	1	9	0	0	0	2	2	0	0
DEM	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
ORD	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
COL	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
AUX	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
NEG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
PRON	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

of-domain text. The Lang-8[5] offers a social network service (SNS) of multi-language essay-correction for foreign language learners. The service has over 400,000 registered members at present and supports 98 languages, facilitating multilingual communication on the web. When learners enter a passage in their target language, native speakers of the language correct the errors for them on the web. This service can provide a huge corpus of language learners' essays, a useful resource for language teachers and learners [33].

[5] http://www.lang-8.com

3.3.2 Methods

Learning-Based Error Type Classifcaton

We use a machine-learning method for the error type classification experiment according to [45]. 13,152 instances are extracted from the NAIST Goyo corpus for an in-domain experiment. We propose an approach for automatic error type classification using error annotated corpus by a machine learning method.

Problem Setting

Figure 3.2 shows a work flow of automatic error type classification. For training text, the instances are already annotated and are not needed for alignment. The incorrect part *(x)*, its correction *(y)* and their error type *(t)* are extracted as *(x, y, t)* from an annotated sentence. Below is an example sentence (for *"I understand English."*):

- (私は) 英語 ***を/が** 分かる
- (watashi wa[6]) Eigo *wo/ga wakaru
- (*I*, a subject-marker) understand English.
- Use of Postposition (P)

Learners are supposed to use "ga" instead of "wo". "Eigo (*English*)" is an object so it is supposed to use with an object-marker. Although the object-marker is in most of cases "wo", learners are likely to use "wo" instead of "ga" as the object marker, which is normally as a subject-marker. "ga" is used also functions as an object-marker when it is used with certain verbs such as "wakaru (*to understand*)." This choice of postposition is one of the difficult grammar aspects for LJ.

The incorrect part *(x)* is "wo" , the correction *(y)* is "ga" and its error type *(t)* is "Postposition (P)" in this case. After that, the contextual information is extracted such as surface strings from a window size of 1 to 3 and POS information. We refer the window size of 1 at both sides as W1, the window size of 2 as W2, and the window size of 3 as W3. The dependency are also extracted as features to train the Maximum Entropy classifier [7].

[6] "wastashi" is a subject and "wa" is a topic(or subject)-marker. In Japanese, subject is often omitted, when the subject is clear to the listener.

[7] `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

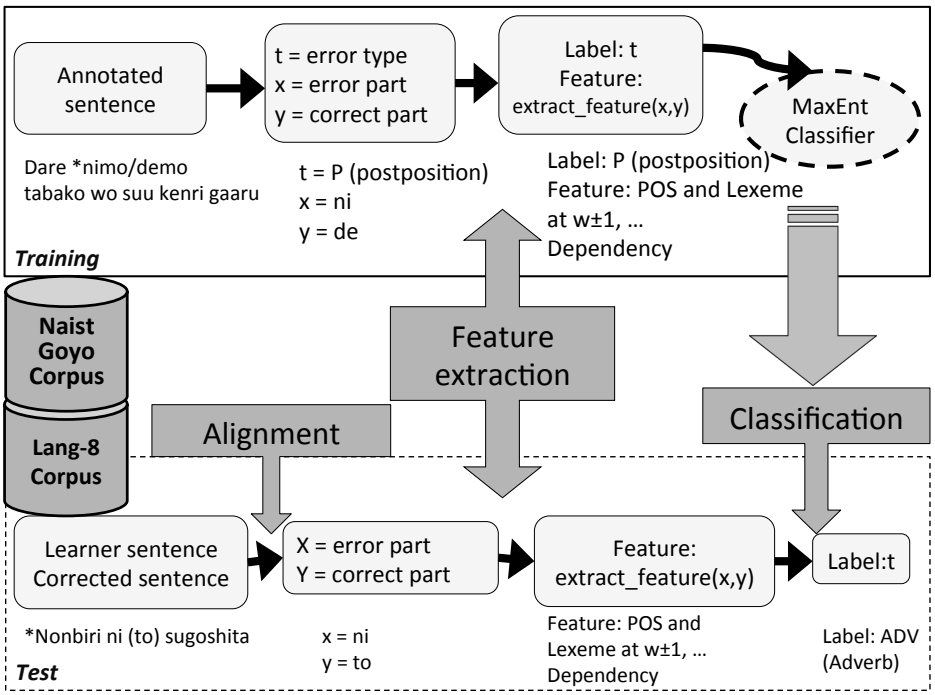


Figure 3.2. Work Flow of Machine-Learning Based Error Type Classification

As for the out-of-domain experiment, the test texts are taken from the Lang-8 corpus, we aligned the incorrect with correct sentences by the dynamic programming method [14][8] and then extracted the contextual information and dependencies as features. The trained model then predicts the error type of an instance from the test texts.

Features

Table 3.4 shows a summary of the features taken from "Eigo *wo/ga wakaru (*I understand English*)". First, we set a baseline where the features are the contextual information of both correct and incorrect instances. We assign POS from the UniDic-2.1.1

[8] `https://github.com/tkyf/jpair`

dictionary using the MeCab-0.994[9] for the contextual information. Next, we added dependency features of both the error and the corrected strings as the extended features using the CaboCha-0.664[10]. We used the surface words and the POS information from W1 to W3 on either side of the target incorrect strings and their corrections.

We considered that the dependency information is effective since the Japanese teachers focussed on it in the human error type classification experiment. Japanese sentence has a free word order, because particles, a group of the postpositions, carry the case to indicate the function of the preceding words. Thus, the words, which depend on each other, are sometimes placed in far or freely in a sentence. Even in such cases, the dependency features work for such a distant relationship in a sentence, in addition to the most immediate context features. When the incorrect string is a verb, the dependency information is taken from the noun depending on a verb. In a sentence of "watashi wa ringo wo *tabetta/tabeta (*I ate an apple*)", given "tabetta (*ate*)" as a conjugational error, "ringo (*an apple*)", "watashi (*I*)", "wa (a topic-marker)" and "wo (an object-marker)" are chosen as dependency features. If the error strings are adjectives, a dependent head noun and a particle are taken as dependency features. For example, consider the sentence "atama ga ii (literally "*head is good*", which meaning "*smart*")". If "ii (*good*)" is a target incorrect string, both "atama (*head*)" and "ga (a subject-marker)" are taken as dependency features. If the incorrect string is a particle, then its head noun and verb depending on the head noun are chosen. From the example of "Eigo *wo/ga wakaru (*I understand English*)", "Eigo (*English*)" and "wakaru (*to understand*)" are taken as features. If the error parts are not dividable and include plural clauses, all bag-of-words in those clauses are taken as features.

Edit Distance Scores

We add edit distance scores in order to see the distance between error parts and correct parts as another extended feature. Extraction is carried out for replacement pairs matching by dynamic programming between correct parts and error parts [14]. The strings are an "addition" error type when they appear in error parts, not correct parts. Similarly, the strings are an "omission" error type when they appear in correct parts, not

[9] `http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html`
[10] `http://code.google.com/p/cabocha/`

Table 3.4. Features of "Eigo *wo/ga wakaru"

Features	Incorrect / Correct samples
Baseline features	
Error part	wo
Correct part	ga
POS, error part (root form)	particle, wo
POS, correct part (root form)	particle, ga
Words (root form), POS at W±1	Eigo (*English*), noun, wakaru (*to understand*), verb
Words (root form), POS at W±2	BOS, EOS
Words (root form), POS at W±3	BOS, EOS
Dependency	eigo (*English*), wakaru (*to understand*)
Extended features (including the baseline)	
Edit distance	1
Substitute probability	0.074 (correct), 0.05 (error)

error parts. The strings are "replacement" when they are replaced to another strings. We count these changes as score 1.

Substitution Probability

We further add substitution probability extracted from the correct and error part paris from the Lang-8 corpusas another extended feature. Lang-8 is a large collection of language learners' writing, which also has error sentences and their correction. This probability is given only by the large body of language learners' corpus such as Lang-8.

Substitution probability is calculated by the frequency of the pairs of the correct and error strings. The number of these pairs extracted from Lang-8 are 796,403. As one example, the error probability is calculated as follows when the error part is "wo" and the correct part is "ga".

$$P(correction = \text{“}ga\text{”} | error = \text{“}wo\text{”}) = \frac{P(correction = \text{“}ga\text{”}, error = \text{“}wo\text{”})}{P(error = \text{“}wo\text{”})} \quad (3.1)$$

The correction probability is calculated as follows.

$$P(error = \text{“}wo\text{”} | correction = \text{“}ga\text{”}) = \frac{P(error = \text{“}wo\text{”}, correction = \text{“}ga\text{”})}{P(correction = \text{“}ga\text{”})} \quad (3.2)$$

3.4. Results

3.4.1 Assessment Measure

Recall (R) indicates the proportion of correctly classified sentences to the sentences belonging to each error type. Precision (P) indicates the correctly classified sentences in proportion to the sentences classified by the system. F-measure (F) shows the harmonic mean of precision and recall. Accuracy (A) shows the proportion of correctly classified and unclassified sentences to all sentences, which is the proportion of true positives to true negatives over all sentences.

$$R = \frac{Correctly\ Classified\ Sentences}{Sentences\ in\ Each\ Error\ Type} \times 100 \quad (3.3)$$

$$P = \frac{Correctly\ Classified\ Sentences}{Classified\ Sentences\ by\ System} \times 100 \quad (3.4)$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.5)$$

$$A = \frac{True\ positives\ and\ True\ negatives}{All\ Sentences} \times 100 \quad (3.6)$$

3.4.2 Experiment with the Tree-Structured Tag Set

We constructed a three-step classification with the tree-structured tag set for both texts. The first step was a multi-class classification of the 3 categories: "OM", "AD" or "Replacement". Second, within "Replacement", binary classification was performed between Grammatical and Lexical groups. Thirdly, we conducted a multi-class classification within two groups, the Grammatical and the Lexical group, with the result of the previous steps.

The last column shows macro average. As the table 3.5 shows, the experiment with the tree-structured tag set increased results approximately by about 13 points from the baseline and the extended features (Table 3.5). The error types of "SEM", "NOT", "V" and "DEM" increased more than 0.8 points in F score with the tree-structured tag set, while only the error type of "P", "OM" and "AD" increased with the flat-structured tag set. That score implies the classification can be conducted with the practical performance.

Unlike other error types, the "OM" and "AD" scores decreased with the tree-structured tag set. It is caused when the classification between "OM", "AD" and "Replacement" is conducted in the first step. The number of "Replacement" is far more than other two groups consisting 10,534 instances compared to 1,441 of "OM" and 1,177 of "AD".

3.4.3 Experiment with Extended Features on the NAIST Goyo corpus

We performed a 10-fold cross validation experiment in the NAIST Goyo corpus. Table 3.6 shows F-measure scores of the baseline feature (BL.), those of the extended features (Ext.) and those of the tree-structured tag set. Baseline features (BL.) are the surface words and POS information from W1 to W3, dependency information and the use of the tree-structured tag set. Extended features are 1) BL. + edit distance score (edit), 2) BL.+ substitution probability (sub.). ALL in Table 3.6 includes all of these extended features. The lower most line in Table 3.6 shows macro average. The Table 3.7 shows micro average of the result of 10-fold cross validation in the NAIST Goyo corpus. The first line shows the result of muliti-class classification between "OM", "AD" and "Replacement". The second and third line shows the result of muliti-class

Table 3.5. Experiment with and without the Tree-Structured Tag Set (10 c.v.) (F score)

	Flat-structure			Tree-structure			
	W1	W2	W3	W1	W2	W3	#
P	96.4	96.3	96.0	98.6	98.3	98.5	3,351
SEM	67.8	65.8	65.3	84.1	83.9	83.2	2,546
NOT	75.5	74.1	72.8	79.4	78.5	77.7	1,830
OM	95.9	95.4	95.5	89.1	89.2	89.1	1,441
V	67.0	66.7	65.4	87.8	87.5	87.1	1,348
AD	89.2	89.7	89.2	83.0	82.0	80.6	1,177
STL	56.9	55.7	53.1	60.1	63.1	61.9	328
NOM	61.6	57.0	58.0	71.5	70.9	71.4	300
CONJ	45.7	42.3	42.7	64.6	63.2	61.1	196
ADJ	44.2	40.4	35.3	75.1	73.0	68.5	149
DEM	61.7	56.1	55.6	83.8	83.6	82.2	137
ORD	17.7	25.6	25.0	42.6	48.7	49.3	121
COL	12.4	11.9	11.1	18.2	16.0	17.4	113
AUX	19.8	31.7	30.4	38.7	42.4	30.6	49
NEG	6.1	11.1	11.1	30.0	29.6	25.6	26
ADV	20.0	16.2	22.4	38.3	36.9	13.3	24
PRON	5.4	7.1	10.0	8.3	14.6	0.0	16
Ave.	49.6	49.6	49.3	62.0	**62.4**	58.7	13,152

classification within “Grammatical” and “Lexical”.

In Table 3.6, Edit distance score improved the performance by 4.1 points in macro average (W2) and substitutional score improved by 3.4 points. The combination of these two features raised 6 points.

Table 3.7 also shows that edit distance score improved the performance by 0.9 points from the BL. and substitutional score improved by 1.2 points in the first step. ALL shows the increase of 1.2 points. Among “Grammatical” group, edit distance score improved the performance by 0.7 points and substitutional score improved by 1.1 points in the first step. ALL shows the increase of 1.2 points. Among “Lexical” group, edit distance score improved the performance by 7.0 points and substitutional score improved by 0.5 points in the first step. ALL shows the increase of 7.4 points.

Table 3.6. Results of 10-fold Cross Validation in NAIST Goyo Corpus (Macro ave. (F-score))

	BL.			BL. + edit			BL. + sub.			ALL			
	W1	W2	W3	W1	W2	W3	W1	W2	W3	W1	W2	W3	#
P	98.6	98.3	98.5	98.7	98.5	98.4	**98.8**	98.6	98.6	98.7	98.6	98.6	3,351
SEM	84.1	83.9	83.2	89.9	89.7	89.5	84.4	84.0	83.1	**90.2**	89.9	89.5	2,546
NOT	79.4	78.5	77.7	88.1	88.0	87.6	79.7	79.1	78.5	**88.5**	88.0	87.8	1,830
OM	89.1	89.2	89.1	90.6	90.2	89.6	91.1	90.4	89.7	**91.2**	90.6	90.5	1,441
V	87.8	87.5	87.1	88.4	88.0	88.3	88.9	88.6	88.3	88.8	**89.6**	89.0	1,348
AD	83.0	82.0	80.6	86.4	86.9	86.1	87.4	86.3	86.4	**87.6**	87.3	87.4	1,177
STL	60.1	63.1	61.9	67.8	65.5	65.7	67.7	68.7	65.7	68.0	**70.8**	67.2	328
NOM	71.5	70.9	71.4	74.3	73.0	72.8	73.9	73.5	73.0	72.5	**75.3**	72.5	300
CONJ	64.6	63.2	61.1	63.5	60.6	63.3	67.1	61.7	62.6	**66.2**	60.2	61.6	196
ADJ	75.1	73.0	68.5	79.8	76.4	74.8	**82.5**	76.8	73.5	79.4	78.5	77.8	149
DEM	83.8	83.6	82.2	83.7	85.8	79.5	85.8	81.7	85.2	**87.3**	85.5	82.3	137
ORD	42.6	48.7	49.3	53.5	54.0	**57.1**	53.2	40.7	45.0	53.5	46.3	46.1	121
COL	18.2	16.0	17.4	15.5	20.2	18.9	**27.6**	23.9	18.7	21.8	20.9	18.4	113
AUX	38.7	42.4	30.6	44.9	48.9	32.5	41.2	53.7	38.4	**55.6**	46.1	42.7	49
NEG	30.0	29.6	25.6	31.8	23.7	15.0	25.0	15.0	0.0	**45.8**	26.0	16.7	26
ADV	38.3	36.9	13.3	12.5	56.7	48.0	38.9	55.6	29.2	56.7	44.4	**73.3**	24
PRON	8.3	14.6	0.0	11.2	23.8	6.3	25.0	18.8	**25.7**	11.1	16.7	0.0	16
Ave.	62.0	62.4	58.7	63.6	66.5	63.1	65.8	64.5	61.3	**68.4**	65.6	64.8	13,152

3.4.4 Experiment in the Lang-8 corpus

We performed the classification on the Lang-8 corpus to see if our classifier is applicable to out-of-domain texts as well. The tree-structured tag set is also the most effective to the out-of-domain classification and improves from 36.1 to 48.5; from 34.8 to 47.8 with W2; from 33.9 to 47.6 with W3 in F-measure as in Table 3.8.

Similarly with the NAIST Goyo corpus experiment, the local information provided with W1 offers high score. However, there are a couple of the error types that show their better classification performance with W2 and W3 in the experiment with the tree-structured tag set. They are ones with W2 such as "Verb (V)", "Postposition (P) ", "Nominalization (NOM)" and "Adjective (ADJ)"; with W3 such as "Connecting (CONJ)" , "Adverb (ADV)" and "Unnecessary (AD)".

Although it is also the case that the lacking of available number of the instances af-

Table 3.7. Results of 10-fold Cross Validation in NAIST Goyo Corpus (Micro ave. (F-score))

	BL.			BL. + edit			BL. + sub.			ALL		
	W1	W2	W3	W1	W2	W3	W1	W2	W3	W1	W2	W3
OM/AD/RE	94.75	94.62	94.40	95.66	95.63	95.42	95.95	95.57	95.48	**95.98**	95.79	95.77
Grammatical	89.88	89.66	89.28	90.57	90.29	90.06	91.02	90.56	90.27	90.90	**91.04**	90.42
Lexical	81.08	80.76	79.92	88.05	87.84	87.59	81.59	81.16	80.10	**88.46**	87.99	87.71

fects the classification performance such as "Collocation (COL)" and "Pronoun (PRON)", the score of "Adverb (ADV)" is lower and the score of "Demonstrative (DEM)" and "Missing (OM)" is higher despite more number of instances. "Pronoun (PRON)" has only 16 instances in the NAIST Goyo corpus for the training, which can be the cause to make the classification difficult in the Lang-8 corpus as well.

Even though the performance with the Lang-8 corpus texts are not as well as that with the NAIST Goyo corpus, the tree-structure tag set has improved the classification performance with the out-of-domain texts.

3.5. Discussion

We discuss the characteristics of the successful and the unsuccessful instances. There are a large number of instances under "OM", "AD", "SEM", "NOT", "P[11]" and "V" and they occupy 91 points of all instances. Thus, we focus on these error types.

The experiment with both extended features improved its performance. As for the experiment with edit distance score, "SEM", "NOT" and "AD" show a improved result. "SEM" includes error in the use of Kanji. With the substitution probability onto the edit distance, the performance of "SEM" is improved further (44.8 points among successful instances). Since "SEM" error instances are of great variety, even a web-based corpus such as Lang-8 is used, there is a possibility that the pattern never appears. Even if this case happens, edit distance still have an effect in the classification

[11] "P" has the most number of instances and also the best result without adding the extended features. Thus, we exclude it from discussion.

Table 3.8. Results in the Lang-8 corpus with the tree-structured tag set (F-score)

	NGC (BL.)			L8 (BL.)			L8 (Tree)			
	W±1	W±2	W±3	W±1	W±2	W±3	W±1	W±2	W±3	#
P	98.6	98.3	98.5	78.7	77.4	76.8	92.2	92.4	91.2	86
SEM	84.1	83.9	83.2	35.2	33.3	34.0	59.2	55.8	56.3	103
NOT	79.4	78.5	77.7	43.4	43.3	44.3	62.8	58.4	60.1	71
OM	89.1	89.2	89.1	68.8	69.6	68.8	65.0	66.7	65.8	56
V	87.8	87.5	87.1	48.8	45.7	49.7	70.8	72.3	71.1	113
AD	83.0	82.0	80.6	64.3	64.2	62.7	61.5	62.4	64.3	62
STL	60.1	63.1	61.9	36.6	33.3	31.2	49.6	47.5	47.9	56
NOM	71.5	70.9	71.4	19.8	16.5	16.8	57.1	57.5	54.5	82
CONJ	64.6	63.2	61.1	46.2	28.0	29.5	44.0	49.6	55.9	73
ADJ	75.1	73.0	68.5	47.2	38.1	43.6	51.4	57.5	51.4	77
DEM	83.8	83.6	82.2	57.5	63.8	50.0	83.2	79.6	76.9	59
ORD	42.6	48.7	49.3	28.6	21.6	16.7	38.5	33.3	32.0	33
COL	18.2	16.0	17.4	0.0	0.0	0.0	13.4	13.3	13.3	29
AUX	38.7	42.4	30.6	10.7	13.1	16.1	32.0	22.2	22.2	50
NEG	30.0	29.6	25.6	18.9	32.8	21.4	25.9	18.9	19.2	53
ADV	38.3	36.9	13.3	8.7	11.1	14.1	17.4	25.0	26.1	64
PRON	8.3	14.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23
ALL	62.0	**62.4**	58.7	36.1	34.8	33.9	**48.5**	47.8	47.6	1,090

performance. A correctly classified example of “SEM” is below:

> 《SEM》このたばこというものはどうして人々の*必用 → 必需品になっているのかがわからない.
>
> kono tabako toiu mono wa doushite hitobito no *hitsuyou → hitsujyuhin ni natteirunoka ga wakaranai.
>
> I don’t know why cigarettes are necessary for people.

The performance of “SEM” is also improved when hiragana or katakana are changed into kanji and vice versa (41.8 points among successful instances). The writing of

Japanese language learners contains full of inappropriate use of hiragana and katakana instead of using of kanji. Edit distance differs largely in the number of characters between hiragana/katakana and kanji, which is effective to the performance.

> 《SEM》いろいろな飾りが大好きだからたばこを買う代わりに欲しがっている*飾り物 → アクセサリーを買った方がいい.
>
> iroirona kazari ga daisuki dakara tabako wo kau kawari ni hoshigatteiru *kazarimono → akusesarii wo katta hou ga ii.
>
> As she likes many kinds of ornaments, you should buy her accessary that she wants.

The performance of "NOT" is also improved with the edit distance feature when hiragana is changed into kanji (55.9 points among successful instances).

> 《NOT》家の外と中を*そうじ → 掃除しました.
>
> ie no soto to naka wo *souji → souji shimashita.
>
> I cleaned up outside and inside of my house.

The performance of "AD" is also improved with the edit distance feature especially when the longer strings are added unnecessarily.

> 《AD》私はその中で*いろいろな食べ物 → ϕ一番好きな物はレマンです.
>
> watashi wa sono naka de*iroirona tabemono → ϕ ichiban sukina mono wa reman desu.
>
> I like lemons best among that.

Next, as for the experiment with substitution probability, “OM” and “V” show a improved result. An example of correctly classified as “V” is below:

> 《V》個人的には，軽い生活磁器よりも韓国の魂が感じ*る → られる
> 非生活磁器が気に入りました.
>
> kojinteki niwa karui seikatsujiki yorimo kankoku no tamashii ga kanji*ru → rareru
> hiseikatsujiki ga kiniirimashita.
>
> Personally, I preferred non dairy-life porcelain to dairy-life porcelain because I felt Korean sprit.
>
> error substitution probability = 0.046, correct substitution probability = 0.475

This example is classified mistakingly as “STL” in the experiment with edit distance feature, while it is classified correctly as “V” with the substitution probability. Since “V” and “STL” include the errors at the end of the sentence, they are classified mistakingly each other. The pattern that “ru” is changed into “rareru” appears often in Lang-8 and its substitution probability is calculated.

“STL”, as explained in Appendix A, is an error whether the sentence ends consistently either “desu/masu” style or “dearu” style. These corrections follows the ones in Taiyaku DB, which is made by professional Japanese teachers. The teachers checked “ru” as an error and corrected to “masu”. Thus, this error-correction pattern (and vice versa) is found repeatedly and is considered a unique pattern for “STL” error type.

Next, although “OM” and “AD” can be classified clearly because “OM” has a missing string in the error part and “AD” has a missing string in the correction part. If the missing and adding part is clearly judged as “P” or other error types, we classified them as “P” or others prior to “OM” or “AD” (see explanation in Appendix A). Thus, even if the error or correction part contains no strings, it is difficult to classify that it is “OM” or “AD” because it contains “P” or other error type. Since “P” occupies 25 points of all instances, the missing or adding strings in “P” also affect the classification of “OM” and “AD”.

The possible cause that the classification performance of "AD" is lower than "OM" can be explained that according to unsuccessful instances "AD" has 10 points more instances that can be considered as "P" or other error types than "OM". This is because these extended features could work in the difference between error and correct strings which has led to distinguish from other error types.

Entirely, even if the longer the context extends, it does not lead to the performance improvement. Swanson and Yamngil [45] also used the context at the window size of 1 from the error part. The longer context does not also contribute in the task for the writing of English language learners. This could be because the morphological analysis in language learners' writing already falls into error. Robust morphological analyzer for the text with grammatical or even spelling errors is also expected in the future.

3.6. Conclusion

This chapter presented an approach to classifying error types in the writing of learners of Japanese language with an error-annotated corpus. We performed a classification experiment with the NAIST Goyo corpus for an in-domain experiment and the Lang-8 corpus for an inter-corpus experiment. The classification results showed the improved performance with the tree-structured classification model, combined with the context features and the dependency information. The tree-structured tag set is highly effective to the classification performance and improves even with an out-of-domain corpus.

However, we further consider the difference of domain in the aim for the improvement on an out-of-domain corpus.

After the experiments, we discussed the characteristics of failure in the classification task. First, it is found that frequently appearing strings classified as an error and its correction are effective for the performance of classification. One way of taking advantages of this influence is to increase the size of corpus and to find more frequently appearing patterns in it. Then, language learners use hiragana even where they are expected to use kanji, which causes a wrong word segmentation and alignment. The word segmentation alignment technology is developing in the machine translation field. The more advance those applications, the better the classification becomes. In addition, since our texts are the essays that language learners wrote, there is a possibility that

the texts include unknown words. In order to reduce this influence, other features are considered to be included. We also consider the nature of LJ's texts to make error type classification better.

The learner corpora have been constructed and the size of them have been growing, however, they are difficult to use directly for linguistic or educational research because they contain both correct and incorrect sentences. Classifying these widely varying incorrect texts into meaningful groups according to their error types benefits language researchers by shedding a linguistic light on how people learn a second language. It also provides learners and teachers feedback on why the errors are made.

Chapter 4

Japanese Particle Error Detection

4.1. Introduction

The goal of this chapter is to automatically detect case particle errors in Japanese learners' writing by featuring at the local contextual cues around a target particle. Automatic error detection is an important task for helping to enrich learner corpora with error information.

A supervised approach is proposed here to learn which particle is most appropriate in a given context by representing the context as a vector populated by features referring to its syntactic characteristics. Support Vector Machines (SVMs) are used with preprocessing methods to identify appropriate particle usage in a newspaper corpus. First, related work is discussed on Japanese case particle error detection. Second, we illustrate the particle identification on a newspaper corpus and error detection experiments on a learner corpus. Finally, we discuss the results and the incorrect usages of particle "wo" that are retrieved in the experiment.

4.2. Previous Research on Automatic Error Detection

Error detection research have been conducted for the purposes such as to check the performance of a machine translation system [44] and to check for errors in Japanese learners' writing [21, 36]. Imaeda et.al [21] proposed a method based on grammar rules and semantic analysis with a case frame dictionary for detection and correction

for LJ's writing.

Nampo et.al [36] also examined detection and correction method for all of the Japanese postpositions (not limited to case particles) by using the clause information in a sentence. They separated a sentence into clauses and used surface forms, POS for each word in the target clause, the dependent clause and the clauses neighboring the target clause. For example, in a sentence "watashi-wa ringo-mo mikan-mo sukidesu" (I like both apples and oranges.) if the clause "mikan-mo" (oranges, too) is considered as a target clause, then the particle or POS information of the neighboring clause, "ringo-mo" (apples, too) are used as features. They reported a recall of 84 points and a precision of 64 points for detection, and a recall of 14 points and a precision of 78 points for correction. However, Nampo et.al [36] conducted evaluation on only 84 selected sentences from learners' essays, which may be too small-scale to present an accurate assessment of its effectiveness. As Chodorow [6] mention, it is difficult to build a model of incorrect usages in learners' writing. Thus, in this study, a model is created for detecting appropriate usages in a newspaper text corpus and accentuate incorrect usages of Japanese particles with the newspaper text model.

4.3. Automatic Detection of Japanese Case Particles on a Newspaper Corpus

4.3.1 Appropriate Case Particle Model

Particle errors are frequent in LJ's writing and are likely to result in misunderstanding of a sentence. A newspaper corpus is used for creating a model that diagnoses correct use of case particles. Table 4.1 shows the number of all 8 case particles ("ga", "wo", "ni", "de", "to", "he", "yori" and "kara") appearing in Mainichi-shimbun Japanese newspapers for half a year. As the figure shows, "wo" is the most frequent, followed by "ni", "ga", "de", "to" and so forth. Five most frequently occurring case particles are selected and a model is trained to choose a proper usage of a particle. A binary classification is used to decide between one case particle and the other particles such as the particle "ga" vs. the others, the particle "wo" vs. the others, and so forth.

Table 4.1. The Number of Occurrences of Case Particles

wo (を)	434,570
ni (に)	408,906
ga (が)	353,139
de (で)	269,232
to (と)	255,583
kara (から)	66,112
he (へ)	21,288
yori (より)	6,510

4.3.2 Experimental Setup: Language Model

An N-gram model is used for sentence features to identify a correct language model. An N-gram language model is based on the idea that a word (or a letter) is affected by neighboring words or letters. As Firth (1957) famously states: "you shall know a word by the company it keeps (p.11), "the collocating words are a key to learn which particle is most appropriate in a given context. If the combination of the words appears often, there is a strong collocation relation among those words. "N" indicates the number of a word such as N=1, 2, 3 and these are referred to as uni-gram, bi-gram and tri-gram models, respectively [28](cf. Table 4.2). An N-gram model can predict the "N" th item by using the N-1 th item as a condition. For example, the bi-gram language model is based on the probability of two words (or letters) occurring together; the occurrence of a word depends on one previous item in a certain context, which represents how strongly the two items collocate. A word-level N-gram model is often used for error detection with the machine learning method, SVMs.

4.3.3 Machine Learning Method

SVMs, which are methods for categorization, is used to train the machine learning models used in the experiments[1]. Training examples are labeled positive or negative and tagged with features: in this experiment, positives are sentences using one particle such as "ga" and negatives are sentences not using "ga". The features are used to map each piece of data into a multi-dimensional space. If the features are similar, they are

[1] http://chasen.org/~taku/software/TinySVM/

Table 4.2. Example of N-gram Collocation

1 (uni)-gram	a	あ	sky	空
2 (bi)-gram	ab	あい	sky is	空は
3 (tri)-gram	abc	あいう	sky is blue	空は青

Table 4.3. Training & Test Set

training	test
10,000	1,000
50,000	5,000
100,000	10,000
200,000	20,000

mapped closely with each other. In this way the two different classes are separated into two groups. SVMs maximize the differences between positive and negative examples; that is, the mathematical modeling is optimized to learn what the difference is between these two groups.

4.3.4 Data

The data was from half-a-year's worth of articles from Mainichi-shimbun, a Japanese newspaper in 2003, which consists of about 20 million words. The data was separated into training data and test data with a ratio of ten to one. In this experiment, 10,000 instances are chosen (one instance consists of one particle with surrounding word information) for the training data and 1,000 for the test data: 50,000 for the training data and 5,000 for the test data and so on.

4.3.5 Procedure

Figure 4.1 shows the flow of the case particle detection experiment. Sentences are first morphologically analyzed by MeCab-0.994. Then, the surface form and POS information was extracted from the words surrounding the target particles. SVMs were trained to create a language model to diagnose whether one of the case particle among

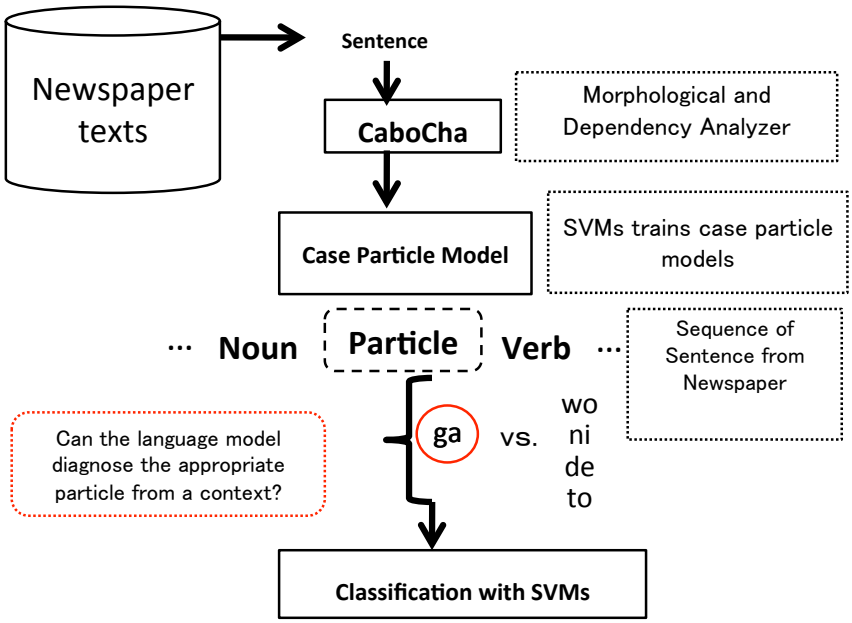


Figure 4.1. Flow of Case Particle Identification Experiment

5 particles is appropriate in a given context. Each of the classifiers was tested with test data to confirm how accurate the classifier was with the metric below.

Because SVMs optimize the difference between two groups, it is advisable to use the features that highlight the divergence between the groups. The following features are used for SVMs: 1) surface forms of words, 2) POS information within a window of ± 3 words from the target case particle. In Figure 4.2, the target case particle is "wo" and the surface forms of the tokens such as "nado (such as)", "no (of)", "katsudou (activity)" are considered before "wo", and "sasaeru (to support)", "supootaa (supporters)", "wo (particle)" are taken as the features after "wo." In addition, POS information such as "nado", "no" ,"katsudou" is considered before "wo" and "sasaeru", "sapootaa", "wo" are considered after "wo." The dependency shows that the verb "sasaeru" determines that "wo" is required to use that verb in this sentence.

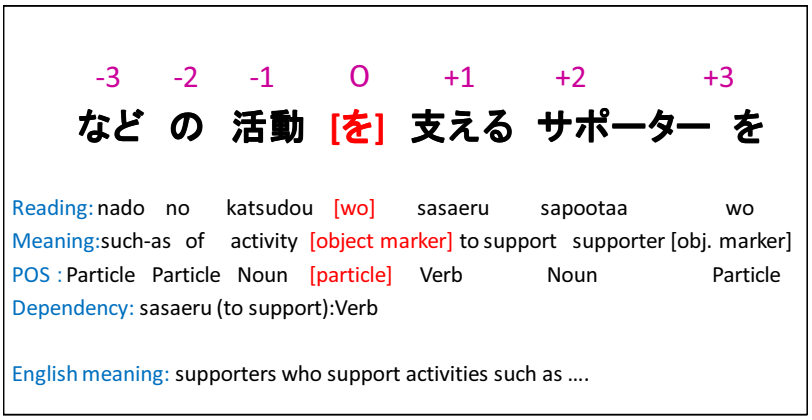


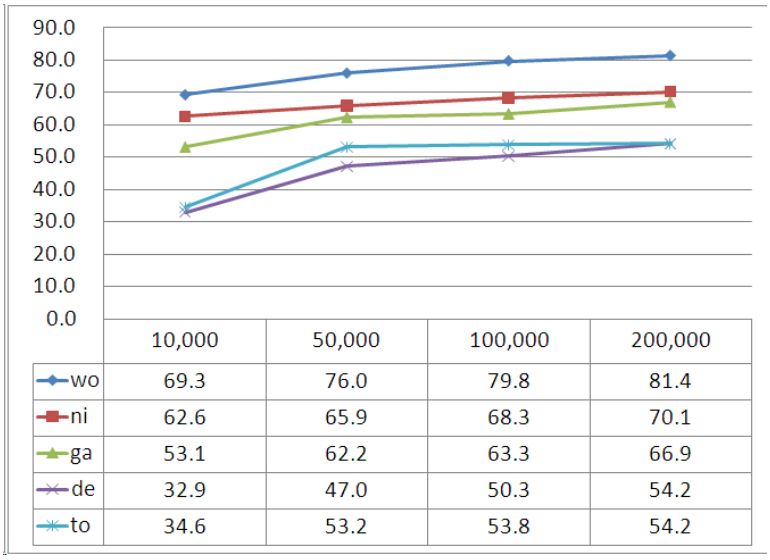
Figure 4.2. Overview of Input Features

4.3.6 Results

The result differs according to each particle. The graph in Figure 4.3 shows the result of the experiment. The horizontal line represents size of the corpus and vertical line represents F score. The object marker "wo" had the best score of 81.4 points, 70.1 points for "ni", 66.9 points for "ga" and 54.2 points for "de" and "to." The "wo" is more easily detectable than the other particles, including "to" or "de", which have lower scores. The reason for low scores for "de" and "to" may reside in the fact that the models for those particles were not trained as much because they were less frequently used in the text than "ni" or "wo." This result depends largely on the frequency of particles distribution in the corpus for this experiment.

4.4. Automatic Detection of Japanese Case Particles on a Learner Corpus

A following experiment is performed to see how the learned model is used to detect wrong usages of particles in a learner corpus, the NAIST Goyo corpus. The "wo" model gave the best score among other particles, so the "wo" particle model is used to detect wrong usages of "wo" in learners' writing. The test data consist of sentences with correct and incorrect usages of "wo." In this experiment, 100 and 200 instances are extracted from the learner corpus. In the 100-instance test set, there are 27 wrong



	10,000	50,000	100,000	200,000
wo	69.3	76.0	79.8	81.4
ni	62.6	65.9	68.3	70.1
ga	53.1	62.2	63.3	66.9
de	32.9	47.0	50.3	54.2
to	34.6	53.2	53.8	54.2

Figure 4.3. Result of Case Particle Experiment

Table 4.4. Result of Error Detection Experiment of “wo”

	100	200
Precision	92.6	95.2
Recall	34.3	37.6
F score	50.0	53.9

usages and 73 correct usages of “wo.” In the 200-instance test set, there are 43 wrong usages and 157 correct usages of “wo.” The result shows 92.6 points for precision and 34.3 points for recall with the former test set, and 95.2 points for precision and 37.6 points recall for the latter.

Here, recall is the ratio of correctly retrieved wrong usages of “wo” to all incorrect instances, while precision is the ratio of correctly retrieved wrong usages to all retrieved instances. One of the reasons for low recall results derives from the variations of Japanese writing, that is, the learners tend to use hiragana, Japanese phonetic characters, whereas kanji, or Chinese characters. The training sentences are extracted from Japanese newspaper articles where kanji characters are always used when appropriate. Thus, when the learners use hiragana instead of kanji, the model classifies them as

wrong usages since it has never seen such usages even though they are not errors. On the other hand, high precision shows that the model has high performance in detecting incorrect usages of particles. Below are the examples of sentences that were retrieved as incorrect usages of “wo.”

- けんこうのため*を → と　しんじる
  - kenkou no tame *wo→ to shinjiru
  - (I) believe in it is for the health.
- あなた　が　このきせい *を → に　さんどうする
  - anata ga kono kisei *wo→ ni sandousuru
  - you agree with this regulation.
- でも たばこ *を → に さわることが...
  - demo tabako *wo→ ni sawaru koto ga
  - But to touch a tabacco...

4.5. Conclusion

In this chapter, we proposed an approach for detecting appropriate usage models of Japanese case particles in order to create an automatic error detection system for LJ’s writing. The experiment resulted in different performance scores according to the kind of case particle, and the case particle “wo” had the most significant result among all case particles. This finding may depend heavily on the number of frequency that the particle appears in a text; “wo” being the most frequently occurring particle in the corpus for this experiment. Future studies will also examine how the choice of different features affects the results and how much the appropriate model approach can help automatic case particle error detection. Additionally, the number of each particle is equalized to minimize the disadvantages of the less-frequent particles for the next experiment.

Chapter 5

Conclusion

This chapter summarizes the main contributions and observations and suggests future work. Language teachers and linguistics have the desire to use learner corpora for research to analyze characteristics of learners' language. The corpora are worth of research because they can describe the actual use of learners' language and show qualitative and quantitative analysis. However, they consist not only learners' correct usages of the target language but also learners' errors. To this end, semi-automatic method of error annotation or any mode of corpora facilitation is necessary since tag annotation currently depends on manual annotation; however, it takes time and cost. In order to utilize learner corpora for linguistic and research study, we have done several attempts for pre-processings in this thesis.

In Chapter 2, we have described constructing a learner corpus by investigating error types and error tags. We also analyzed a particle usage in the the NAIST Goyo corpus. We have learned that a particle omission type is the most frequent error and especially "no" and "wa" are the most difficult among all particles. Since only particle errors are investigated this time, other types of error is investigated such as "Verb" error or "Word choice" error in the future work.

In Chapter 3, we have described error type classification in the NAIST Goyo corpus. The experiment resulted 80 points in precision altogether, which leads to realize an automatic error tag annotation application. We also conducted an inter-corpurs experiment on Lang-8. The experiment on the out-of-domain corpus shows a lower accuracy than the in-domain text by 14.9 points. For the future work, we will consider the difference of the domain for the out-of-domain corpus to be classified with

the model trained by other corpora.

In Chapter 4, we have described a Japanese particle detection task. We created classification model for the usage of “wo” with a newspaper corpus. The classification model of “wo” was applied to a learner corpus in order to distinguish an error sentence from a correct sentence. In the 100-instance test set, the result showed 50.0 points in F scores and in the 200-instance test set, it showed 53.9 points. For the future work, we will consider oversampling to equalize of unbalanced number of each particle in order to reduce the influence of lack of instances.

Acknowledgements

I gratefully acknowledge the support and generosity of Mr. Xi YangYang who permitted to use the Lang-8 corpus, without which the present study could not have been completed. I am also thankful for the reviewers of Natural Language Processing journal for beneficial comments.

Appendix

A. Error Type Description

Postposition (P) includes omission, addition of a postposition. It also includes a choice of a wrong postposition or compound particle.

Word Choice (SEM) includes inappropriate word selection due to not considering context. For example, both "**bubun**" and "**ichibu**" in Japanese as in Table 3.1 can be rendered "some" in English, but "some people" is translated into "**ichibu no hito**" and "*__bubun jin__" is not acceptable.

Spelling (NOT) includes wrong use of the three types of Japanese characters: Hiragana, Katakana and Kanji.

Missing (OM) indicates that the sentence has an element missing. However, if the missing element can also be classified as "Postposition (P)" , "Adverb (V)" and so forth, those categories are prioritized.

Verb (V) covers a wide range of types, such as errors in verb conjugation, transitive or intransitive verb form choice, errors with the passive voice, with tense/aspect and so forth.

Unnecessary (AD) indicates that unnecessary words or expressions are written in a sentence, making it ungrammatical or unnatural.

Inappropriate register (STL) includes the wrong choice of sentence ending. A Japanese essay text must be consistent, distinguishing throughout the "desu/masu" level from a written style using, for example, "da/dearu". Inconsistency of sentence ending is one of the major errors and is often seen in LJ's writing.

Nominalization (NOM) in Japanese (as in "to watch/watching" in English) requires choosing "no" or "koto," depending on the context, a rule which confuses learners. "*Shumi wa eiga wo miru **no** desu" contains an error; "Shumi wa eiga wo miru **koto** desu (I enjoy **watching** a movie)" is correct. On the other hand, "Tori ga tobu **no** wo mimashita (I saw a bird **flying** in the sky)" uses, but not "*Tori ga tobu **koto** wo mimashita".

Connecting (CONJ) is an error in conjunction use (corresponding to "and","then", "because" and etc in English). There are two forms for "because" in Japanese, "node" and "kara". The former is rather used in a written form and the latter is in a spoken

form.

Adjective (ADJ) includes a wrong choice of adjectives and a conjugational error. A Japanese adjective conjugates in its combinations with a verb, adverb or noun that follows it. The adjective suffix "-i" is used before nouns. The other suffix "-kute" is used before verbs.

Demonstrative (DEM) includes errors in the use of "ko", "so" or "a" which fall into three categories according to distance from the participants in a dialogue. These distinctions are not found in the native languages of many LJ and they often err here.

Word order (ORD) is also important; given the case particles in Japanese, word order is more flexible than in English, but not all of the order. As in Table 3.1, "shichigatsu **yori** (from July)" is correct, not the English-like word order "***yori** shichigatsu".

The **Collocation (COL)** category consists of a mistaken combination of noun-particle-verb, while "Word Choice (SEM)" takes a wrong choice of a word.

The **use of da (AUX)** follows grammatical rules unique to Japanese. Japanese complex sentences require that the subordinate clause should end in the copula "da," as in "Ano hito wa kirei **da** to omoimasu (I think **that** that girl is pretty)". The copula "da" becomes "desu" at the end of a polite sentence. The difficulty of this distinction leads to errors like "*Ano hito wa kirei **desu** to omoimasu", where "da" is replaced by "desu".

Negation (NEG) includes expressions by negating verbs and the use of negational conjunction "nakute" and "naide". The "nakute" means "because I do not" and "naide" means "without". "Ie ni irare **nakute** soto e ikimashita (I went out **because I just could not** stay in the house.)"; "*Ie ni irare **naide** soto e ikimashita" is not used. "naide" is used in, for example, "Kasa wo mota**naide** ie wo demashita. (I left home **without** bringing an umblera.)".

Some **adverbs (ADV)** are used with either "ni" or "to" particles in Japanese, differentiated according to the preceding word, though they are completely interchangeable in some contexts. "Nonbiri (slowly)" collocates with "to", so that "*Nonbiri **ni** sugoshita" is unacceptable; "Nonbiri **to** sugoshita (I spend a day **at leisure**)" is the acceptable form.

For the **Pronoun (PRON)** category, both "*Kare**tachi**" meaning of "they" is not acceptable but "Kare**ra**" is.

We created multiple instances out of sentences that contain multiple errors. When

an instance was classified into more than one category, we used the most likely tag since those instances exist only 3 points of the whole text.

B. 76 Error Types

Table 5.1. 76 Error Types

Description	Sample and Correction	English Translation	Tag
P	*Eigo ***wo/ga** wakaru	I can understand English	p
SEM	{ ***bubun jin / ichibu no hito** }	**some people**	sem
NOT (Kanji)	hito ni{***ai (合い) / ai (会い)**} masu	I **meet** a person	not/kg
NOT (kana)	nen{***pa / pai**} no hito	the **elderly** people	not/hg
NOT (symbols)	kinou{***ϕ / ,**} tomodachi ni atta	Yesterday **,** I met my friend	not
OM	Nobu {***ϕ / toiu**} resutoran ni ikimashita	I went to a restaurant **called** Nobu	om
V (conjugation)	tegami wo {***kaki / kaka**}nai	I **will not write** a letter.	v/jug
V (volitional form)	watashi no taiken wo {***tsutae / tsutaeyou**} to omou	I **will tell** my experience	v/othr/vol(*)
V (volitional form)	ashita yan san ga {***tsukou / tsukudarou**} to omou	I think that yan san **will arrive** tomorrow	v/vol/othr
V (causative form)	kono seikou wa sekaikakkoku wo **{*odoroita / odorokaseta}**	This sucess **made the world surprise**	v/othr/cs
V (causative form)	yasashii kotoba ga ningenkankei wo {***iji-sasete / iji-shite**}kureru	Kind remarks **maintains human relation**	v/cs/othr
V (causative passive)	nihongo no kyouiku ni{***kyouhaku-saserare**} {**kyousei-saserare**}mashita	We were **forced to take** Japanese education	v/csp/othr
V (causative passive)	watashi wa haha ni osara wo {***aratta / arawaserareta**}	My mother **made me to wash** dishes	v/othr/csp
V (passive form)	honkon wa tyugoku he {***utsurare/ utsusare**}mashita	The capital of China **was relocated** from Nanjin to Beijin	v/othr/psv
V (passive form)	nihon no shinryaku niyorui {***henkan-suru koto ni naru / henkan-sareru**}	Hongkong will**return** to China damage from Japanese occupation	v/othr/psv
V (passive form)	1949 nen tyugoku no shuto wa nankin karai {***utsurare/ utsusare**}mashita	The capital of China **relocated** from Nanjin to Beijin	v/psv/intran
V (passive form)	doubutsu wo{***shinarete/ shinasete**}shimatta	I **let an animal die**	v/psv/cs

(*) The tag "v/vol/othr" indicates that there is a "vol" (volitional form) error and it is changed to "othr" (other words) inside "v" (verb) category. The tag "v/othr/vol" is an opposite case of above.

Description	Sample and Correction	English Translation	Tag
V (potential form)	watashi no yume wo motto hayaku {***utsurare/ utsusare**}mashita	My dream will **be realized** quicker	v/psv/intran
V (potential form)	{***soudan-shite kurerareremasen**} **/ soudan-shite-kuremsen**}	(He) **does not ask** me	v/pot/othr
V (potential form)	nihongo wo chigau shiten de {***mireru / mirareru**}	(We) **can take a different aspect** on Japanese	v/pot/othr
V (tran/intran verb)	oto ga{***ageru / agaru**}	Sound **goes up**	v/tran/intran
V (tran/intran verb)	shokuji wo{***owatte / owete**}	After (we) **finish eating**	v/intran/tran
V (imperative verb)	tabako wo {***suuna / suwanaide**}kudasai	**Don't smoke**	v/imp/othr
V (imperative verb)	{***suwanai / suuna**} to meirei-shimasen	**I don't order not to smoke**	v/othr/imp
V (tense)	wakai koro hon ga{***daisukina / daisukidatta**} node	**When I was young, I liked a book**	v/t_prs/pt
V (tense)	nihon ni {***kita/kuru**} maeni	**Before I came to Japan, ...**	v/t_pt/prs
V (aspect)	yopparatte {***tanoshimi wo shiteiru / tanoshindeiru**}	**(He) enjoys drinking**	v/a_othr/teiru
V (aspect)	nichijyou tsukawareteiru kotoba mo {***koushin-shiteiru / koushin-suru**}keikou mo aru	It seems that the everyday words are being **renewed**	v/a_teiru/othr
V (aspect)	watashi wa nihon ni {***suimasu / sundeimasu**}	**I live in Japan**	v/a_sta/teiru
V (aspect)	nomu inryou mo {***hanbai-saretearu / hanbai-sareteiru**}	**Beverages are also sold**	v/a_tearu/teiru
V (aspect)	hatsuon ga kawaruto, imi mo {***chigakunaru / chigattekuru**}	When the pronunciation changes, the meaning also **changes**	v/a_othr/tekuru
V (aspect)	konoyo wo {***ikiru / ikiteiku**}	**I survive in this life**	v/a_othr/teiku
V (aspect)	yoku onyomi to kunyomi wo mazete {***hanashimasu / hanashiteshimaimasu**}	I often **speak** by mixing onyomi and kunyomi	v/a_teiru/othr
V (aspect)	rekishi wo {***mimashou / mitemimashou**}	**Let's take a look at** the history	v/a_othr/temiru
V (aspect)	paatii no mae ni jyuusu wo {***kaimasu / katteokimasu**}	**Let's take a look at the history**	v/a_othr/temiru
V (aspect)	shukudai wo {***shimasu/shitekara**} terebi wo mimasu	I watch TV **after finishing** homework	v/a_othr/tekara
V (give&receive)	daigaku ga ryuugakusei no tameni osewa wo {***shiteiru/shitekudasaru**}to kikimashita	The university **takes care of** international students	v/othr/gr
V (give&receive)	ane ga watashitachi wo jinja he {***tsureteikimashita/tsureteittekuremashita**}	My sister **took us** to the shrine	v/othr/gr
V (give&receive)	bideo wo{***kaitai desuga {/katte rokuga shitai**}	**I want to buy a video and record Japanese reality**	v/te
V (sino-Japanese)	{***benkyou-jin/benkyou-suru hito**}	**Those who study**	v/sa
V (compound verbs)	{***tsuzukete motteimasu/mochitsuzuketeimasu**}	**(We) keep possessing**	v/othr/cmp
V (iru/aru)	hito ga{***aru/iru**}	**There is** a person	v/iru/aru
V (iru/aru)	tsukue ga{***iru/aru**}	**There is** a desk	v/aru/iru
V (polite form)	watashi wa nihon wo{***gozonji desu/shitteimasu**}	I **know** Japan	v/pol/othr
V (polite form)	sensei ga{***kuru/irassharu**}	My teacher **comes** here	v/othr/pol

Description	Sample and Correction	English Translation	Tag
AD	{ *__tenki ga / ϕ__samukute... }	**The weather** is so cold...	ad
STL	Totemo taihen {__*ne / desu__}	It is very hard	stl
NOM	shumi wa eiga wo miru{*__no / koto__} desu	I enjoy **watching** a movie	nom
CONJ	{ *__Soshitemo / Soshite__} pet to asobimasu	**And then,** I played with my pet	conj
CONJ (conjunction particles)	hashi wo{__*tsukutte/tsukuru node__} orihime to hikoboshi wa hashi de aemasu	**As the bridge will be made** orihime and hikoboshi can meet there	conj/part
CONJ	{ *__Soshitemo / Soshite__} pet to asobimasu	**And then,** I played with my pet	conj
ADJ	boku wa {__*huto–kute / huto–i__}hito desukara	I am a **fat** person	adj
ADJ (adj. conjunction)	{__*tanoshii deshita/tanoshikatta desu__}	**I had a fun**	adj/jug
DEM	{ *__Asoko / Soko__}de tomodachi ni aimashita.	I met a friend **there**	dem
ORD	{ *__yori shichigatsu / shichigatsu yori__ }	**From** July	ord
COL	Shiken {__*ni sankashimashita / wo ukemashita__}	I **took a test**	col
AUX	Anohito wa kirei{__*desu/da__}to omoimasu	I think **that** the girl is pretty	aux/da
AUX (auxiliary)	ashita wa ame{__*deshou/darou__} to omoimasu	I think that it **will rain** tomorrow	aux
NEG (NEG)	Ie ni irare {*__naide / nakute__} soto he ikimashita	I went out **because I did not** want to stay at home	neg
ADV	Nonbiri {*__ni / to__} sugoshita	I spend a day **at leisure**	adv
PRON	{ *__Karetachi / Karera__ }	**they /them**	pron
Nouns	nichiyoubi wa gakkou ga {*__yasumimasu / yasumidesu__}	**Sunday is a school holiday.**	noun
Nouns (formal noun)	benkyou dekiru {*__tameni/youni__}	**In order to study**	noun/keishiki
Noun modification	{ *__Iwayuru / ϕ__} kokusaikankei	**so-called** international relationship	nmod/noun
Noun modification (verb)	keizai__{hattatsuno/ga hatten-shiteiru}__ senshinkoku	**Advanced countries of** **high economy growth**	nmod/verb
Noun modification (adjective)	shougakkou notoki onaji__{na/ϕ}__ elementary school	**We went to the same**	nmod/adj
Modality	Sensei wa beiryugakushite imawa tyugoku ni ikitai{*__desukedo / soudesu__}	my teacher wants to study in China.	md
Phrases	{ *__Otoko teare Onna teare__ Otoko deare Onna deare}	Regardless of men or women.	ph
Whole alteration	{ *__fukai kankeishite tabako wo suu__} { __tabako wo suu noniwa hukai kankei ga aru__}	**There is a deep relation between smoking** **(and getting rid of stress).**	ful

Description	Sample and Correction	English Translation	Tag
Miscellaneous	See Table 5.1		–
Miscellaneous (no desu)	watasshi wa nittyu kankei no shigoto wo **{*nda/ϕ}**to omoimasu	**Advanced countries of high economy growth**	no
Miscellaneous (conditional form)	kyouto ni ikunara shinkansen ga benri desu	**If you go to kyoto,** shinkansen is a convenient transportation	cond
Miscellaneous (toki)	tai no omatsuri ga kuru ***toki/to**, tai jin to gaikoku jin ga au	**When you go to the tai festival,** Tai people meet foreigners	toki
Miscellaneous (tai)	resutoran wo {***hanareru /hanaretai**} to kangaeteiru	**I want to leave the restaurant.**	tai
Miscellaneous (organization)	{ ***jyunbi no koto desu**} **jyunbi wa, tsugi no youni shimasu. mazu,... tsugi ni ...**	**Preparation will be made as following.** First,... Next,...	org
Miscellaneous (comparison)	tai wa nihon hodo { ***atsui/atsukunai**}	**Thai is not as hot as Japan** First,... Next,...	compa
Miscellaneous (numbering)	{ ***ikutsuka no hi/ikunichi ka**}	**Some days**	num
Miscellaneous (hosii)	watashi to issho no toki wa tabako wo { ***suwanai hou ga ii/suwanaide hoshii**}	**I hope you don't smoke** if you are around me	hosii

Table 5.2. Further Tag Definition in 76 Error Types

Tag	Definition
"p"	particle
"v"	verb
"jug"	conjugation
"cs"	causative
"csp"	causative passive
"psv"	passive
"intran"	intransitive
"tran"	transitive
"pot"	potential form
"ra"	ra is omitted
"imp"	imperative form
"t_prs"	tense present
"t_pt"	tense past
"a_teiru"	aspect teiru
"sta"	state
"a_tearu"	aspect tearu
"a_tekuru"	aspect tekuru
"a_teiku"	aspect teiru
"a_tesimau"	aspect tesimau
"a_temiru"	aspect temiru
"a_teoku"	aspect teoku
"a_tekara"	aspect tekara
"gr"	giving & receiving
"cmp"	compound verbs
"pol"	polite form
"conj/part"	conjunction particle
"nmod"	noun modification
"md"	modality
"cond"	conditional form
"org"	organization
"compa"	comparison
"num"	number

References

[1] I. Asada. An analysis of acquisition data of homonymous adverbs in corpora of jsl writing. *Project Report of Graduate School of Social and Cultural Sciences at Kumamoto University 2007 No. 7*, pages 79–96, 2007.

[2] I. Asada. Word order of adverbs by chinese learners of japanese. *roject Report of Graduate School of Social and Cultural Sciences at Kumamoto University 2008 No. 8*, pages 37–58, 2008.

[3] C. Brockett, W.B. Dolan, and M. Gamon. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 249–256, Sydney, Australia, 2006.

[4] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[5] H. Chang. Native and non-native speakers' usages of "i (watashi wa)" in essays. *Gakkou Kyouikugaku Kenkyu Ronshu No.22*, pages 23 –35, 2010.

[6] M. Chodorow and C. Leacock. An unsupervised method for detecting grammatical errors. In *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 140–147, Seattle, U.S.A, 2000.

[7] M. Chodorow, J. Tetreault, and N-R. Han. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL–SIGSEM Workshop on Prepositions*, pages 45–50, Prague, Czech Public, 2007.

[8] E. Dagneaux, S. Denness, and S. Granger. Computer-aided error analysis. 26:163–174, 1998.

[9] E. Dagneaux, S. Denness, and S. Granger. Computer-aided error analysis. *System*, 26:163–174, 1998.

[10] R. De Felice and S.G. Pulman. Automatically acquiring models of prepositional use. In *Proceedings of the 4th ACL–SIGSEM Workshop on Prepositions*, pages 45–50, Prague, Czech Public, 2007.

[11] R. De Felice and S.G. Pulman. A classifier-based approach to preposition and determiner error correction in L2. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 169–176, Manchester, U.K., 2008.

[12] N.C. Ellis. Constructions, chunking, and connectionism:the emergence of second language structure. In C. Doughty and M. Long, editors, *The handbook of second language acquisition*. Blackwell, 2003.

[13] E. Fitzpatrick and M.S. Seegmiller. The montclair electronic language database project. In U. Connor and T.A. Upton, editors, *Applied Corpus Linguistics. A Multidimensional Perspective*, pages 223–237. Rodopi, 2004.

[14] T. Fujino, T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto. Word segmentation for automatic error correction in the Japanese language learners' essays. In *Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing*, pages 26–29, 2012.

[15] M. Gamon, J. Gao, C. Brockett, A. Klementiev, W.B. Dolan, D. Belenko, and L. Vanderwende. Using contextual speller techniques and language modelling for ESL error correction. In *Proceedings of the 3rd International Joint Conference on Computational Linguistics (IJCNLP 2008)*, pages 449–456, Hyderabad, India, 2008.

[16] S. Granger. Uses of tenses by advanced EFL learners: evidence from an error-tagged computer corpus. 26:163–174, 1999.

[17] S. Granger. Error-tagged learner corpus and CALL:A promising synergy. *The Computer Assisted Language Instruction Consortium (CALICO)*, 20:465–480, 2003.

[18] N. R. Han, M. Chodorow, and C. Leacock. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129, 2006.

[19] Y. Ichikawa. *A dictionary of Japanese Language Learners' Errors*. Bonjinsha, 1997.

[20] Y. Ichikawa. *A Dictionary of Japanese Language Learners' Errors* II. Bonjinsha, 2000.

[21] K. Imaeda, A. Kawai, Y. Ishikawa, R. Nagata, and F. Masui. Error detection and correction of case particles in Japanese learner's composition. In *Proceedings of the Information Processing Society of Japan SIG*, pages 39–46, 2003.

[22] K. Imamura, K. Saito, K. Sadamitsu, and H. Nishikawa. Grammar error correction using pseudo-error sentences and domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 388–392, 2012.

[23] E. Izumi, K. Uchimoto, and H. Isahawa. *The* NICT JLE *Corpus : a guideline for error annotation ver.1.1.* ALC, 2004.

[24] C. James. *Errors in Language Learning and Use: Exploring Error Analysis*. Addison Wesley, 1998.

[25] H. Y. Kageyama. *Shin Hajimeteno Nihongo Kyouiku 1 – Nihongo kyouikuno Kisochishiki –*. ASK, Tokyo, 2004.

[26] O. Kamata and H. Yamauchi. KY corpus versition 1.1. Report, Vocabulary Acquisition Study Group, 1999.

[27] J. Lee, G. Lin, Y. Miyaoka, and H. Shibasaki. Creation of Japanese language learners' corpus with application of the natural language processing. In *Proceedings of the Spring Meeting of the Society of Japanese Language and Linguistics in 2012*, 2012.

[28] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, U.K., 1999.

[29] T. Masuoka and Y. Takubo. *Kiso Nihongo Bumpo*. Kuroshio, 1992.

[30] M. Matsuda, A. Mori, K. Kanamura, and H. Goto. Error and language transfer in noun phrases of learners of japanese: analysis based on the essay data in japanese language written by people of 7 countries in asia. *Foreign Students Education 11*, pages 45–53, 2006.

[31] E. Mays, F.J. Damerau, and R.L. Mercer. Context based spelling correction. *Information Processing and Management*, 23(5):517–522, 1991.

[32] F. Mine. Learning process of naide and nakute from the point of view of the development of language processing. In *Seventh International Conference on Practical Linguistics of Japanese (ICPLJ7)*, pages 92–93, San Francisco, U.S.A, 2011.

[33] T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto. Mining revision log of language learning SNS for automated Japanese error correction. *Journal of the Japanese Society for Artificial Intelliigence No.4*, pages 420–432, 2013.

[34] Yoshinaga N. The difference between the two Japanese negative forms of predicates : "naide" and "nakute". *Sonoda Women's University Papers No.47*, pages 133–140, 2013.

[35] R. Nagata, T. Wakana, A. Kawai, K. Morihiro, F. Masui, and N. Isu. Recognizing errors in English writing based on the mass count distinction. *The Institute of Electronics, Information and Communication Engineers (IEICE), Transactions on Information and Systems*, J89-D(8):1777–1790, 2006.

[36] R. Nampo, H. Ototake, and K. Araki. Automatic error detection and correction of Japanese particles using features within bunsetsu. In *Proceedings of the Information Processing Society of Japan SIG*, pages 107–112, 2007.

[37] D. Nicholls. The Cambridge Learner Corpus–error coding and analysis for lexicography and ELT. In Archer et al., editor, *Proceedings of the Corpus Linguistics 2003 Conference(CL2003)*, pages 572–581. 2003.

[38] M. Ohki, H. Oyama, S. Kitauchi, T. Suenaga, and Y. Matsumoto. Error detection in the system manual texts by non-Japanese native speakers. In *Proceedings of The 17th Annual Meeting of The Association for Natural Language Processing*, pages 1047–1050, 2011.

[39] M. Oso, M. Sugiura, Y. Ichikawa, M. Okumura, S. Komori, H. Shirai, N. Takizawa, and T. Sotoike. A learners' corpus of Japanese compositions: Digitalizing and sharing the data. Report, University of Nagoya, 1998.

[40] H. Oyama, K. Mamoru, and Y. Matsumoto. Towards automatic error type classification of japanese language learners' writings. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 163–172, Taipei, Taiwan, 2013.

[41] H. Oyama and Y. Matsumoto. Automatic error detection method for Japanese particles. *Polyglossia Vol.18*, pages 55–63, 2010.

[42] M. Shimizu, F. Du, and M. Dantsuji. A project to construct the Chinese learners' parallel corpus of Japanese and develop DUT corpus linguistics tools. In *Proceedings of the 6th International Symposium on Applied Linguistics and Language Leaning*, pages 585–595, 2005.

[43] G. Sun, X. Liu, G. Cong, M. Zhou, Z. Xiong, J. Lee, and C.Y. Lin. Detecting erroneous sentences using automatically mined sequential patterns. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 81–88, Prague, Czech Public, 2007.

[44] H. Suzuki and K. Toutanova. Learning to predict case makers in Japanese. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1049–1056, Sydney, Australia, 2006.

[45] B. Swanson and E. Yamangil. Correction detection and error type selection as an ESL educational aid. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 357–361, Montreal, Canada, 2012.

[46] H. Teramura. Examples of error sentences for the Japanese language learners–conjunctions and adverbs–. Technical report, Osaka University and The Natonal Iinstitute of Japanese Language, 1990.

[47] H. Terashima. Application of corpus in japanese education – data-driven learning and discussion of its practices. *Polyglossia*, pages 91–103, 2013.

[48] J. Tetreault and M. Chodorow. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on*

Computational Linguistics (COLING 2008), pages 865–872, Manchester, U.K., 2008.

[49] G. Wang. A study on the expressions in which Chinese learners of Japanese frequently make mistakes. *Hokuriku University Papers 27*, pages 115–122, 2003.

[50] A. Wilcox-O'Hearn, G. Hirst, and A. Budanitsky. Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. In A.Gelbukh, editor, *Proceedings of 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)(Lecture Notes in Computer Science Vol.4919)*, pages 605–616, Berlin, 2008. Springer.

List of Publications

- Refereed Journal Papers

  1. H. Oyama, M. Komachi and Y. Matsumoto. Hierarchical Annotation and Automatic Error-Type Classification of Japanese Language Learners' Writing. *Journal of Natural Language Processing, Vol.23 No.2*, 2016 Mar.

- Non-Refereed Papers

  1. H. Oyama. Automatic Error Detection Method for Japanese Language Learners' Support System. *Language Issues Vol.16*, The Prefectural University of Kumamoto. pp.87–99, 2010 Feb.
  2. H. Oyama. Automatic Error Detection Method for Japanese Particles. Ritsumeikan Asia Pacific University *Polyglossia Vol.18*, pp.55–63, 2010 Feb.
  3. H. Oyama. Designing an Error Tag Set for the Corpus of Japanese Language Learners. *Japanese Language Literature Research 54*, pp.102–114, 2009.

- Refereed Conference Papers

  1. H. Oyama, M. Komachi and Y. Matsumoto. Towards automatic error type classification of Japanese language learners' writings. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages pp.163–172, Taipei, Taiwan, 2013.
  2. H. Oyama. Error pattern of writing of Japanese learners of multi-nationality by multivariate analysis. In *Proceedings of the Society for Teaching Japanese as a Foreign Language*, pp.159–164, 2010 Oct.
  3. H. Oyama. Automatic error detection method for Japanese particles–Basic computational experiment to assist the writing test–. In *Proceedings of the Society for Teaching Japanese as a Foreign Language*, pp.169–174, 2010 Oct.
  4. H. Oyama and Y. Matsumoto. Automatic Error Detection Method for Japanese Case Particles in Japanese Language Learners' Writing. *The 6th ICTATLL (ICT in Analysis, Teaching and Learning of Language) International Conference*, pp.235–245, 2010 Sep.

5. H. Oyama and Y. Matsumoto. “Automatic Error Detection for Japanese Case Particle on Balanced Corpus of Contemporary Written Japanese (BCCWJ). In *Proceedings of Open Workshop Satellite Session in The National Institute for Japanese Language and Linguistics (NINJAL)*. 2010 Mar.
6. H. Oyama and Y. Matsumoto. Automatic Error Detection Method for Japanese Language Learners' Support System. In *Proceedings of the Society for Teaching Japanese as a Foreign Language*, pp.169–174, 2008 May.
7. H. Oyama, K. Sakata, Y. Matsumoto and M. Asahara. Construction of an Error Information Tagged Corpus of Japanese Language Learners and Automatic Error Detection. In *Proceeding of Computer Associated Language Instruction Consortium (CALICO)*, 2008 May. (poster presentation)

- Non-Refereed Conference Papers

1. H. Oyama, M. Komachi and Y. Matsumoto. Issues on designing error annotated corpus –in the creation of NAIST GOYO corpus–. In *1st text annotation workshop*, 2012 Aug.
2. H. Oyama and Y. Matsumoto. Error Annotation on Japanese Language Learner Corpora and Automatic Error Detection of Japanese Case Particles. *Basic Research on Corpus Annotation workshop*, 2010 Aug.
3. H. Oyama, M. Komachi and Y. Matsumoto. Automatic Error Detection for Japanese Case Particle for Japanese Language Learners. In *Proceedings of The 13th Annual Meeting of The Association for Natural Language Processing*. pp.787–788, 2007 Mar.