Error Annotation for Corpus of Japanese Learner English

Emi Izumi **Kiyotaka Uchimoto** **Hitoshi Isahara**
National Institute of Information and Communications Technology (NICT),
Computational Linguistics Group
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{emi,uchimoto,isahara}@nict.go.jp

Abstract

In this paper, we discuss how error annotation for learner corpora should be done by explaining the state of the art of error tagging schemes in learner corpus research. Several learner corpora, including the NICT JLE (Japanese Learner English) Corpus that we have compiled are annotated with error tagsets designed by categorizing "likely" errors implied from the existing canonical grammar rules or POS (part-of-speech) system in advance. Such error tagging can help to successfully assess to what extent learners can command the basic language system, especially grammar, but is insufficient for describing learners' communicative competence. To overcome this limitation, we re-examined learner language in the NICT JLE Corpus by focusing on "intelligibility" and "naturalness", and determined how the current error tagset should be revised.

1 Introduction

The growth of corpus research in recent years is evidenced not only by the growing number of new corpora but also by their wider variety. Various "specialized corpora" have recently been created. One of them is the "learner corpus", which is a collection of the language spoken or written by non-native speakers. The primary purpose of learner corpora is to offer Second Language Acquisition (SLA) researchers and language teaching professionals resources for their research. In order to develop a curriculum or pedagogy of language teaching, it would be beneficial to have interlanguage data so that researchers can scientifically describe the characteristics of each developmental stage of their interlanguage. One of the most effective ways of doing this is to analyze learner errors. Some of the existing learner corpora are annotated for errors, and our learner corpus called the "NICT JLE (Japanese Learner English) Corpus" is one of them. This is a two-million-word speech corpus of Japanese learner English. The source of the corpus data is 1,281 audio-recorded speech samples of an English oral proficiency interview test ACTFL-ALC Standard Speaking Test (SST). The advantage of using the SST data as a source is that each speaker's data includes his or her proficiency level based on the SST scoring method, which makes it possible to easily analyze and compare the characteristics of interlanguage of each developmental stage. This is one of the advantages of the NICT JLE corpus that is rarely found in other learner corpora.

Although there are a lot of advantages of error-annotated learner corpora, we found some difficulties in designing an error tagset that covers important features of learner errors. The current version of our error tagset targets morphological, grammatical, and lexical errors, and we found it can help to successfully assess to what extent learners can command the basic language system, especially grammar. However, we also found that it is not sufficient to measure learners' communicative skills. In order to determine how the current error tagset should be extended to cover more communicative aspects

of learner language, we re-examined the learner data in the NICT JLE Corpus by focusing on "intelligibility" and "naturalness".

In this paper, we discuss how error annotation for learner corpora should be designed and actually performed. The remainder of this paper is organized as follows. Section 2 outlines the influence that Error Analysis (EA) in SLA research in the 1970s had on error annotation for learner corpora to enable us to rethink the concept of error annotation. Section 3 provides some examples of learner corpus projects in which error tagging is performed. Section 4 describes the current error tagging scheme for the NICT JLE Corpus. Section 5 examines how we can expand it to make it more useful for measuring learners' communicative skills. Finally, section 6 draws some general conclusions.

2 Error Tagging and EA

The idea of trying to tag errors in learner corpora might come from the notion of EA in SLA research in the 1970s. In order to design an error tagset and to actually perform tagging, we would like to reconfirm the concept of the traditional EA by considering about the definition of learner errors, the importance of analyzing learner errors for describing learner language, the actual EA procedures, and problems, and limitations of EA.

2.1 Definition of Learner Errors

Errors in a second language (L2) are often compared with errors in the first language (L1). According to Ellis (1994), before EA was introduced, L2 errors were often considered as "undesirable forms". On the other hand, errors made by young L1 learners were regarded as the "transitional phase" in L1 acquisition, while errors made by adult L1 speakers were seen just as slips of the tongue. L2 errors were often shoved back into the closet as a negative aspect of learner language.

In EA, errors are treated as evidence that plays an important role in describing learner language. Corder (1981) asserts that talking about learner errors only with terms like "deviant" or "ill-formed" is inappropriate because it leads to learner errors being treated just as superficial deviations. Even if learners produce outputs whose surface structures are well-formed, this is not enough to prove that they have acquired the same language system as that L1 speakers have.

In EA, learner errors are treated as something that proves that learners are in the transitional phase in L2 acquisition in a similar way to treating the language of L1 children. However, it is problematic to assume these two are exactly the same language. In L1 and L2 acquisition, there are certain processes in common, but they have not been scientifically confirmed. Many differences are found between these two. The learner language including errors can be defined as "interlanguage", which lies between L1 and L2 (Selinker, 1972). According to Corder (1981), learner errors are evidence of learning strategies in which learners are "investigating" the system of the new language (L2) by examining to what extent L1 and L2 are similar and how different they are.

2.2 Importance of EA

Analyzing learner errors is important for teachers, researchers, and learners themselves in the following way (Corder, 1981). First, for teachers, errors can give them hints about the extent to which the learners have acquired the language system by that time and what they still have to learn. For SLA researchers, errors can reveal the process by which L2 is acquired and the kinds of strategies or methodology the learners use in that process. Finally, for learners themselves, as stated in 2.1, making errors is one of the most important learning strategies for testing the interlanguage hypothesis that learners have established about L2. In other words, knowing what kinds of errors were made by themselves or by other learners can be "negative evidence (or feedback)" given directly or indirectly to learners that an interlanguage hypothesis is incorrect (Ellis, 1997).

2.3 EA Procedure

In general, the EA procedure can be divided into four stages as shown in Figure 1 (Ellis, 1994).

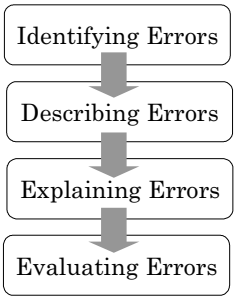


Figure1. EA Procedure.

In the first stage, identifying errors, it is necessary to localize errors by pointing out which letters, words, and phrases, or how sentence structures or word order, are incorrect. In the second stage, identified errors should be described by being linguistically categorized depending on, for example, their POS (part-of-speech), linguistic level (morpheme, syntax, lexis, or discourse), or how they deviate from the correct usage on the surface structure (redundancy, omission, or replacement). Thirdly, "explaining errors" means identifying why those errors occurred. This is a very important task in order to figure out the learners' cognitive stage. Some causes of learner errors have been recognized in common such as errors caused by language transfer, learning and communication strategy-based errors, and the transfer of training and induced errors. Finally, errors are evaluated. This can be done by estimating intelligibility or near-nativeness of erroneous outputs. In other words, "error gravity" is estimated by examining how each error interferes with the intelligibility of the entire outputs.

2.4 Problems and Limitations of Traditional EA

Although it is widely recognized that EA contributes to describing learner language and the improving second language pedagogy, several problems and limitations have been pointed out mainly because a concrete methodology of EA has not been established yet. Most importantly, EA cannot be successful without robust error typology, which is often very difficult to obtain. Since it used to be difficult to collect or access large databases of learner language, a robust error typology that covers almost all error types was not established in traditional EA.

Another criticism against EA is that errors reflect only one side of learner language. A lot of people point out that if a researcher analyzes only errors and neglects what learners can do correctly, he/she will fail to capture the entire picture of learner language. It is time-consuming to count both correct and incorrect usages in learner data, and this must have been quite difficult to do in the past before computing technology was developed.

Furthermore, the real significance of EA cannot be identified without using diachronic data in order to describe learners' developmental stages. The types and frequencies of errors change with each acquisition phase. Without longitudinal data of learner language, it is difficult to obtain a reliable result by EA.

2.5 From EA to Error–coded Learner Corpora

The problems and limitations of traditional EA are mainly due to the deficiency of computing technology and the lack of large databases in early times. However, now that computing technology has advanced, and a lot of learner data is available, it might be possible to perform EA more effectively mainly by annotating errors. Although the basic motivations for error annotation are the same as those of traditional EA, such as describing learner language and improving second-language pedagogy, several new applications of EA might become possible such as the development of a new computer-aided language learning (CALL) environment that can process learners' erroneous input and give feedback automatically.

Degneaux, et al. (1998) call EA based on learner corpora "computer-aided error analysis (CEA)", and expect that the rapid progress of computing technology and learner corpora will be able to solve the problems and overcome the limitations of traditional EA. Surely, thanks to the quantitative database of learner language, we will become able to cover a wider range of learner errors. Advances in computing technology make it possible to perform statistical analysis with quantitative data more easily. However, it must be noted that human researchers still have a lot of work to do in the same manner as in traditional EA, such as establishing an error typology for error tagging

or examining results obtained from CEA carefully.

3 Related Work

There are a few learner corpus projects that implement CEA. For example, in the International Corpus of Learner English (ICLE) project, which was launched by Professor Sylviane Granger at the University of Louvain, Belgium, and has been a "pioneer" in learner corpus research since the early 1990s, they performed error tagging with a custom-designed error tagset (Degneaux, et al., 1996). The grammatical, lexical and pragmatic errors are dealt with in their error tagset and the corrected form is also indicated for each error. We guess that error categorization has been done mainly by translating the basic English grammar or lexical rules into an error ontology to try to cover as many types of errors as possible. The ICLE team currently comprises 17 partners internationally, and the corpus encloses 17 subcorpora of learners of the different mother tongues (Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, Swedish, and so on). A comparison of this data will make possible "contrastive interlanguage analysis (CIA)" proposed by Granger (2002). CIA involves both NN/NNS and NNS/NNS comparisons (NS: native speakers; NNS: non-native speakers), as shown in Figure 2. NS/NNS comparisons might reveal how and why learner language is non-nativelike. NNS/NNS comparisons help researchers to distinguish features shared by several learner populations, which are more likely to be developmental from ones peculiar to a certain NNS group, which may be L1-dependent.

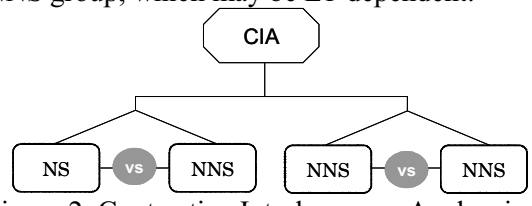


Figure 2. Contrastive Interlanguage Analaysis (Granger, 2002).

Another corpus thathas been error tagged is the "Japanese EFL Learner (JEFLL) Corpus". This corpus, which was created by Professor Yukio Tono at Meikai University in Japan, has three parts: i) the L2 learner corpora which include written (composition) and spoken (picture description) data of Japanese learner English, ii) the L1 corpora consisting of Japanese written texts for the same tasks as those in the first part and Japanese newspaper articles, and iii) the EFL textbook corpus, which is the collection of EFL textbooks used officially at every junior high school in Japan (Tono, 2002). Compared with the ICLE, which has been annotated with the generic error tagset, the error tagging for the JEFFL Corpus focuses on specific types of errors, especially major grammatical morphemes such as articles, plural forms of nouns, and third person singular present forms of verbs, and so on. We assume that those items were selected due to the corpus developer's research interests. Although completeness for covering all errors would be decreased by focusing on a limited number of error types, annotators will be able to perform tagging more stably without being confused among various different types of errors.

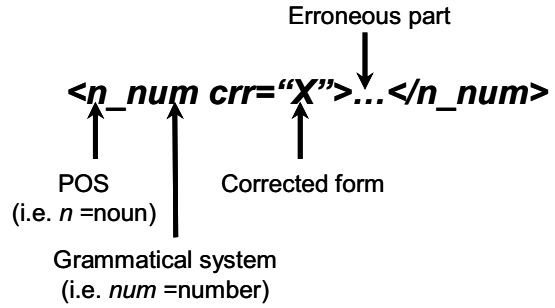
The Cambridge Learners' Corpus (CLC), which has been compiled by Cambridge University Press and Cambridge ESOL (English for Speakers of Other Languages), is also an error-coded learner corpus. It forms part of the Cambridge International Corpus (CIC) and is a large collection of essay writing from learners of English all over the world. This corpus has been utilized for the development of publications by authors and writers in Cambridge University Press and by members of staff at Cambridge ESOL. Over eight million words of the CLC have been error-coded with a Learner Error Coding System devised by Cambridge University Press. In order to make the tagged data as consistent as possible, tagging has been done by only one annotator since it started in 1993. Their error tagset covers 80 types of errors. The annotator chooses an appropriate tag for each error mainly by identifying which POS the error involves and how it deviates from the correct usage (redundancy, omission, or replacement). A corrected form is also indicated for each error.

4 Error Tags in the NICT JLE Corpus

In this section, we introduce the error annotation scheme we used for the NICT JLE Corpus.

We are aware that it is quite difficult to design a consistent error tagset as the learner

errors extend across various linguistic areas, including grammar, lexis, and phoneme, and so on. We designed the original error tagset only for morphological, grammatical, and lexical errors, which are relatively easy to categorize compared with other error types, such as discourse errors and other types of errors related to more communicative aspects of learners' language. As shown in Figure 3, our error tags contain three pieces of information: POS, morphological/grammatical/lexical rules, and a corrected form. For errors that cannot be categorized as they do not belong to any word class, such as the misordering of words, we prepared special tags. The error tagset currently consists of 46 tags (Table 1).



example) *I belong to two baseball <n_num crr= "teams">team</n_num>.*

Figure 3. Structure of an Error Tag and an Example of an Error-tagged Sentence

The tags are based on XML (extensible markup language) syntax. One advantage of using XML is that it can clearly identify the structure of the text and it is also very beneficial when corpus data is utilized for web-based pedagogical tools or databases as a hypertext.

The error tagset was designed based on the concept of the ICLE's error tagging, that is, to deal with as many morphological, grammatical, and lexical errors as possible to have a generic error tagset. However, there are several differences between these two tagsets. For example, in the ICLE, only replacement-type errors are linguistically categorized, and redundant- and omission-type errors are not categorized any more and just called as "word redundant" or "word missing", while in our error tagset, all these three types of errors are linguistically categorized.

Although our error tagset covers major grammatical and lexical errors, annotators often have difficulties to select the most appropriate one for each error in actual tagging process. For example, one erroneous part can often been interpreted as more than one error type, or sometimes multiple errors are overlapping in the same position.

To solve these problems, tagging was done under a few basic principles as follows.

1) Because of the limitation of XML syntax (i.e. Crossing of different tags is not allowed.), each sentence should be corrected in a small unit (word or phrase) and avoid to change a sentence structure unnecessarily.
2) If one phenomenon can be interpreted as more than one error type, select an error type with which an erroneous sentence can be reconstructed into a correct one without changing the sentence structure drastically. In this manner, errors should be annotated as locally as possible, but there is only one exception for prefabricated phrases. For example, if a sentence "*There are lot of books.*" should be corrected into "*There are a lot of books.*", two ways of tagging are possible as shown in a) and b).
   a) *There are <at crr= "a"></at> lot of books.*
   b) *There are <o_lxc crr= "a lot of">lot of</o_lxc> books.*

   In a), just an article "*a*" is added before "lot of", while in b), "lot of" is corrected into "a lot of" as a prefabricated phrase. In this case, b) is preferred.
3) If multiple errors overlap in the same or partly-same position, choose error tags with which an erroneous sentence can be reconstructed into a correct one step by step in order to figure out as many errors as possible. For example, in the case that a sentence "*They are looking monkeys.*" should be corrected into a sentence "*They are watching monkeys.*", two ways of tagging are possible as shown in c) and d).
   c) *They are <v_lxc crr= "watching"> looking</v_lxc> monkeys.*
   d) *They are <v_lxc crr= "watching"> looking<prp_lxc2 crr= "at"> </prp_lxc2></v_lxc> monkeys.*

   In c), "*looking*" is replaced with "*watching*" in one step, while in d), missing of a preposition "*at*" is pointed out first, then, "*looking at*" is replaced

with "*watching*". In our error tagging scheme, d) is more preferred.

Tag	Error category
	NOUN
<n_inf>…</n_inf>	Noun inflection
<n_num>…</n_num>	Noun number
<n_cs>…</n_cs>	Noun case
<n_cnt>…</n_cnt>	Countability of noun
<n_cmp>…</n_cmp>	Complement of noun
<n_lxc>…</n_lxc>	Lexis
	VERB
<v_inf>…</v_inf>	Verb inflection
<v_agr>…</v_agr>	Subject-verb disagreement
<v_fml>…</v_fml>	Verb form
<v_tns>…</v_tns>	Verb tense
<v_asp>…</v_asp>	Verb aspect
<v_vo>…</v_vo>	Verb voice
<v_fin>…</v_fin>	Usage of finite/infinite verb
<v_ng>…</v_ng>	Verb negation
<v_qst>…</v_qst>	Question
<v_cmp>…</v_cmp>	Complement of verb
<v_lxc>…</v_lxc>	Lexis
	MODAL VERB
<mo_lxc>…</mo_lxc>	Lexis
	ADJECTIVE
<aj_inf>…</aj_inf>	Adjective inflection
<aj_us>…</aj_us>	Usage of positive/comparative/superlative of adjective
<aj_num>…</aj_num>	Adjective number
<aj_agr>…</aj_agr>	Number disagreement of adjective
<aj_qnt>…</aj_qnt>	Quantitative adjective
<aj_cmp>…</aj_cmp>	Complement of adjective
<aj_lxc>…</aj_lxc>	Lexis
	ADVERB
<av_inf>…</av_inf>	Adverb inflection
<av_us>…</av_us>	Usage of positive/comparative/superlative of adverb
<av_pst>… </av_pst>	Adverb position
<av_lxc>…</av_lxc>	Lexis
	PREPOSITION
<prp_cmp>…</prp_cmp>	Complement of preposition
<prp_lxc1>…</prp_lxc1>	Normal preposition
<prp_lxc2>…</prp_lxc2>	Dependent preposition
	ARTICLE
<at>…</at>	Article
	PRONOUN
<pn_inf>…</pn_inf>	Pronoun inflection
<pn_agr>…</pn_agr>	Number/sex disagreement of pronoun
<pn_cs>…</pn_cs>	Pronoun case
<pn_lxc>…</pn_lxc>	Lexis
	CONJUNCTION
<con_lxc>…</con_lxc>	Lexis
	RELATIVE PRONOUN
<rel_cs>…</rel_cs>	Case of relative pronoun
<rel_lxc>…</rel_lxc>	Lexis
	INTERROGATIVE
<itr_lxc>…</itr_lxc>	Lexis
	OTHERS
<o_je>…</o_je>	Japanese English
<o_lxc>…</o_lxc>	Collocation
<o_odr>…</o_odr>	Misordering of words
<o_uk>…</o_uk>	Unknown type errors
<o_uit>…</o_uit>	Unintelligible utterance

Table 1. Error Tags for the NICT JLE Corpus.

4.1 Advantages of Current Error Tagset

Error tagging for learner corpora including the NICT JLE Corpus and the other corpora listed in Section 3 is carried out mainly by categorizing "likely" errors implied from the existing canonical grammar rules or POS system in advance. In this sub-section, we examine the advantages of this type of error tagging through research and development done by using these corpora.

Tono (2002) tried to determine the order in which Japanese learners acquire the major English grammatical morphemes using the error tag information in the JEFFL Corpus. Izumi and Isahara (2004) did the same investigation based on the NICT JLE Corpus and found that there was a significant correlation between their sequence and Tono's except for a few differences that we assume arose from the difference in the languga e production medium (written or spoken). Granger (1999) found that French learners of English tended to make verb errors in the simple present and past tenses based on the French component of the ICLE. Izumi et al. (2004) also developed a framework for automated error detection based on machine learning in which the error-tagged data of the NICT JLE Corpus was used as training data. In the experiment, they obtained 50% recall and 76% precision.

Error tagging based on the existing canonical grammar rules or POS system can help to successfully assess to what extent learners can command the basic language system, especially grammar. This can assist people such as teachers who want to improve their grammar teaching method, researchers who want to construct a model of learners' grammatical competence, and learners who are studying for exams with particular emphasis on grammatical accuracy.

5 Future Improvement

Finally, let us explain our plans for future improving and extending error tagging for the NICT JLE Corpus.

5.1 Problems of Current Error Tagset

Although the current error tagging scheme is beneficial in the ways mentioned in 4.1, it cannot be denied that much could be improved to make it useful for teachers and researchers who want to know learners' communicative skills rather than grammatical competence. The same can be said for learners themselves. In the past, English education in Japan mainly focused on developing grammatical competence in the past. However, in recent years, because of the recognition of English as an important communication tool among peoples with different languages or cultures, acquiring communicative competence, especially

production skills, has become the main goal for learners. One of the most important things for acquiring communicative skills might be producing outputs that can be understood properly by others. In other words, for many learners, conveying their messages clearly is often more important than just producing grammatically-correct sentences.

It is necessary to make the current error tagset more useful for measuring learners' communicative competence. To do this, firstly we need to know what kind of learners' outputs can be understood by native speakers and in what cases they fail to convey their messages properly. By doing this, it should become possible to differentiate fatal errors that prevent the entire output from being understood from small errors that do not interfere with understanding.

Another goal of studying English for learners, especially at the advanced level, is to speak like a native speaker. Some learners mind whether their English sounds natural or not to native speakers. In the current error tagging, both obvious errors and expressions that are not errors but are unnatural are treated at the same level. It would be better to differentiate them in the new error annotation scheme.

5.2 Survey for Extending Current Error Tagset

To solve the problems of our current error tagging system discussed in 5.1, we decided to do a survey to:

1) Identify fatal errors and small ones by examining "learners' outputs that can be understood properly by native speakers" and "those that do not make sense to native speakers".
2) Identify unnatural and non-nativelike expressions and examine why they sound unnatural.

We will do this mainly by examining the learner data corrected by a native speaker.

Correction by NS

We asked a native speaker of English to correct raw learner data (15 interviews, 17,068 words, 1,657 sentences) from the NICT JLE Corpus and add one of the following three comments (Table 2) to each part.

Comment 1	It is obviously an error, but does not interfere with understanding.
Comment 2	The meaning of the utterance does not make sense at all.
Comment 3	It is not an error, and the utterance makes sense, but it sounds unnatural.

Table 2. Comments added to each error

The person who did the corrections is a middle-aged British man who has lived in Japan for 14 years. He does not have experience as an English teacher, but used to teach Japanese Linguistics at a British University. Although he is familiar with English spoken by Japanese people because of his long residence in Japan and the knowledge of the Japanese language, we asked him to apply the corrections objectively with considering whether or not each utterance was generally intelligible to native speakers.

Corrected Parts

A total of 959 errors were corrected and 724 of these were labeled with Comment 1, 57 with Comment 2, and 178 with Comment 3, respectively (Table 3).

Comment 1	724
Comment 2	57
Comment 3	178
Total	959

Table 3. Number of Errors Labeled with Each Comment.

In order to examine what kind of differences can be found among errors labeled with these comments, we categorized them into four types (morpheme, grammar, lexis, and discourse) depending on which linguistic level each of them belongs to based on corrected forms and additional comments made by the labeler (Table 4).

	Comment1	Comment2	Comment3	Total
Morpheme	6	0	0	6
Grammar	429	0	52	481
Lexis	286	43	78	407
Discourse	3	14	48	65
Total	724	57	178	959

Table 4. Linguistic Level Involved in Each Error.

As a whole, the most common type was grammar (481), but most of the grammatical errors (or cases of unnaturalness) were labeled with Comment 1, which implies that in most cases, the grammatical errors do not have a fatal influence making the entire output unintelligible. The second-most common type was lexical errors (or cases of unnaturalness) (407). Half of them were labeled with Comment 1, but 23 errors got Comment 2. This means that some

errors can interfere with understanding. Discourse errors accounted for a fraction of a percent of all errors (65). However, compared with other types of errors, the percentage of Comment 2 was the highest (14 out of 65), which means that discourse errors can greatly interfere with the intelligibility of the entire output. The main difference between the discourse errors labeled with Comment 2 and those labeled with Comment 3 was that most of the latter related to collocational expressions, while the former involved non-collocational phrases where learners need to construct a phrase or sentence by combining single words. In the following sections, we examine the characteristics of each type of error (or cases of unnaturalness) in detail.

Comment 1

Half of the Comment 1 errors were grammatical ones. Most of them were local errors such as subject-verb disagreement or article errors. There were 286 lexical errors, but in most cases, they were not very serious, for example lexical confusions among semantically similar vocabulary items.

Comment 2

Most of the Comment 2 errors had something to do with lexis or discourse.

1) Too abrupt literary style (discourse error)
   ex) I've been to the restaurant is first. I took lunch. The curry the restaurant serves is very much, so I was surprised and I'm now a little sleepy.
2) Unclear context (discourse error)
3) Unclear anaphora (pronouns and demonstratives) (discourse error)
4) Mis-selection of vocabulary (lexical error)
5) Omission of an important word (subject, predicate or object) (lexical or syntax error)
   ex) She didn't (*) so much about fashion.
   ex) Last year, I enjoyed living alone, but nowadays, it's a little bit troublesome because I have to (*) all of the things by myself.
6) Japanese English/Direct translation (lexical error)
   ex) bed town (as "bedroom suburbs")
   ex) claim (as "complaint")

Comment 3

There were grammatical, lexical and discourse problems with the parts labeled with Comment 1) Verbose expressions (discourse-level unnaturalness)
ex) T: Can I call you Hanako?
L: Yes, please call me Hanako.
better → Yes, please do.
ex) Three couples are there and they're having dinner.
better → Three couples are having dinner.
ex) I told my friends about this, and my friends agreed with me.
better → I told my friends about this, and they agreed with me.

2) Socio-linguistically inappropriate expressions (discourse/pragmatic-level unnaturalness)
ex) What?
better → I beg your pardon?
ex) Good.
better → I'm fine.

3) Abrupt expressions (discourse/pragmatic-level unnaturalness)
ex) T: Have you been busy lately?
L: No.
better → No, not really.

4) Overstatement (discourse/pragmatic-level unnaturalness)
(In a normal context)
ex) T: How are you?
L: I'm very fine.
better → I'm fine.

5) There are more appropriate words or expressions. (discourse/pragmatic-level of unnaturalness)
ex) To go to high school in the mainland, I went out of the island.
better → … I left the island.

5.3 Limitation of Current Error Annotation Scheme

It is obvious that discourse and some types of lexical errors can often impede the understanding of the entire utterance.

Although our current error tagset does not cover discourse errors, it is still possible to

"just" assign any one of error tags to the erroneous parts shown in 5.2. There are two reasons for this. One is that, in the current error tagging principle, it is possible to replace, add or delete all POS in order to make it possible to "reconstruct" an erroneous sentence into a correct one. The other reason is that since discourse structure is liked to grammatical and lexical selections, it is possible to translate terms for describing discourse into terms for describing grammar or lexis.

However, annotating discourse errors with tags named with grammatical or lexical terms cannot represents the nature of discourse errors. Since discourse errors are often related to intelligibility of learners' outputs, describing those errors with appropriate terms is quite important for making the current error tagset something helpful for measuring learners' communicative competence. We will need to know what kind of discourse errors are made by learners, and classify them to build in the error tagset. Some parts labeled with Comment 3 were also related to discourse-level problems. It would be beneficial to provide learners with feedback such as "Your English sounds unnatural because it's socio-linguistically inappropriate". Therefore, it is also necessary to classify discourse-level unnaturalness in learners language.

5.4 Works for Expansion to New Error Tagset

We decided the basic principles for revising the current error tagset as following.

1) Classify second language discourse errors and building them into a new error tagset.
2) Differentiate unnatural expressions from errors. Information on why it sounds unnatural will also be added.
3) Add information on linguistic level (morpheme, grammar, lexis, and discourse) to each tag.
4) Do a further survey on how we can differentiate errors that interfere with understanding and those that do not, and add information on error gravity to each tag.

Classifying discourse errors will be the most important task in the tagset revision. In several studies, second language discourse has already been discussed (James, 1998), but there is no commonly recognized discourse error typology. Although grammatical and lexical errors can be classified based on the existing canonical grammar rules or POS system, in order to construct the discourse error typology, we will need to do more investigation into "real" samples of learners' discourse errors.

Adding the information on linguistic level (morpheme, grammar, lexis and discourse) to each tag is also important. From the survey, we found that the linguistic level of errors is strongly related to the intelligibility of the entire output. If linguistic level information is added to each error tag, this might help to measure the intelligibility of learners' utterances, that is, learners' communicative competence.

6 Conclusion

In this paper, we discussed how the error annotation scheme for learner corpora should be designed mainly by explaining the current error tagging scheme for the NICT JLE Corpus and its future expansion. Through learner data corrected by a native speaker, we decided to introduce discourse errors into the error annotation in order to cover learners' communicative competence, which cannot be measured with the current error tagging scheme.

References

Corder, P. (1981). *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

Degneaux, E., Denness, S., Granger, S., & Meunier, F. (1996). *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Universite Catholique de Louvain.

Degneaux, E., Denness, S., & Ganger, S. (1998). Computer-aided error analysis, *System*, 26, 163-174.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Ellis, R. (1997). *Second Language Acquisition*. Oxford: Oxford University Press. pp. 47, 67.

Granger, S. (1999). Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus. In Hasselgard, H., & Oksefjell, S. (Eds). *Out of Corpora*. (pp. 191-202). Amsterdam: Rodopi.

Granger, S. (2002) A bird's-eye view of learner corpus research. In Granger, S., Hung, J., and Tyson, P.S. (Eds.). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam: John Benjamins Publishing Company.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The overview of the SST speech corpus of Japanese learner English and evaluation through the experiment on automatic detection of learners' errors. In *Proceedings of Language Resource and Evaluation Conference (LREC) 2004*, Portugal, 1435-1438.

Izumi, E., & Isahara, H. (2004). Investigation into language learners' acquisition order based on the error analysis of the learner corpus. In *Proceedings of Pacific-Asia Conference on Language, Information and Computation (PACLIC) 18 Satellite Workshop on E-Learning*. Tokyo, Japan.

James, C. (1998). *Errors in Language Learning and Use: exploring error analysis*. Essex: Longman.

Selinker, L. (1972). Interlanguage. In Robinett, B. W., & Schachter, J. (Eds.). (1983). *Second Language Learning: Contrastive analysis, error analysis, and related aspects*. (pp. 173-196). Michigan: The University of Michigan Press.

Tono, Y. (2002). *The Role of Learner Corpora in SLA Research and Foreign Language Teaching: The Multiple Comparison Approach*. Unpublished Ph.D. Thesis. Lancaster University, UK.

CLC (Cambridge Learners Corpus)'s Website:

http://uk.cambridge.org/elt/corpus/clc.htm