The NICT JLE Corpus in Well-Formed XML Format: From Analyses of Surface Forms to Functions in Longer Stretches of Discourse

Aika Miura
Hiroshi Sano

1. Introduction

This paper explores the possibilities of expanding the scope of spoken learner corpora from investigations of forms (e.g. lexico-grammatical features) to those of pragmatic functions (e.g. speech act expressions). It focuses on the National Institute of Information and Communications Technology Japanese Learner English Corpus (NICT JLE), which contains the data of approximately 1,200 Japanese English as a foreign language (EFL) learners taking an oral interview test. The corpus comprises annotated transcripts that contain data on the speakers (i.e. interviewers and interviewees), including their level of English proficiency, and extra-linguistic features of the utterances, for example, fillers, pauses, self-corrections, and overlaps between speakers.

Recently, the advent of multi-modal corpora of spoken data has made it possible to analyse not only prosodic features and turn-taking in interactions, but also speakers' gestures, facial expressions, and posture, provided by audio or audio-visual files (Adolphs; Knight and Adolphs; O'Keeffe, Clancy, and Adolphs). Although the NICT JLE Corpus was created before the advent of multi-modal corpora and is composed of only the transcribed data of spoken interactions, marked-up elements annotated in the corpus are highly detailed and informative in the investigation of interactive spoken data of EFL learners. However, the markup used to encode extra-linguistic features in the

NICT JLE Corpus does not correspond to well-formed XML format. For example, there are missing start- and end-tags, especially when the data include intersecting utterances by different speakers. Once the corpus is formatted in valid XML, it will be relatively convenient to investigate the pragmatic functions of the utterances in longer stretches of discourse and to examine the contexts of interactions between the interviewers and interviewees.

The main aim of this paper is to explain the process of converting the NICT JLE Corpus text files into well-formed XML documents and to describe how the XML-converted corpus can be applied to research on pragmatic functions. The paper describes how valid XML documents allow researchers to expand the scope of analysis of the NICT JLE Corpus data. Specifically, the operation of handling the data becomes less constrained, and thus, it is easy to extract specific corpus data according to the researchers' needs (e.g. by segmenting the data by learner groups or interview stages) and to apply additional annotations to the corpus, such as annotations representing pragmatic features.

2. Corpus-based Studies in the Domain of Pragmatics and Interlanguage Pragmatics

With the recent compilation of language corpora, researchers have gained access to a large collection of naturally occurring data (O'Keeffe et al.). However, corpus-based research on the nature of the relationship between linguistic form and function has remained relatively scarce (Adolphs; Knight and Adolphs; O'Keeffe et al.). The study, which aims to explain the disparity between linguistic form and meaning in context, is an investigation in pragmatics. Further, it explores how a speaker's intended meaning to the hearer can be realised in particular linguistic forms. The functions, for example, of speech act expressions and conversational implicatures, are not necessarily equivalent to their surface meaning.

The background to the difficulties of applying a corpus-based approach to pragmatics is as follows. First, not many spoken corpora

have been available for pragmatic analysis, as it is difficult to compile spoken corpora (Adolphs). According to O'Keeffe et al., it takes ten hours to transcribe one hour of talk, which typically consists of approximately 10,000 to 15,000 words. They note that "spoken corpora are few, compared to written corpora, and those that are available may not be designed in such a way that suits the study of pragmatic features" (33). For example, it is necessary to search manually for instances of pragmatic features such as speech acts. Manual tagging for pragmatic features is, of course, time-consuming, yet the ability to interpret a particular function of a speech act expression based on the context is vital (Adolphs). Thus, because it is necessary but not always easy to infer the various contexts of utterances in the corpus (Romero-Trillo), researchers often make their classifications "partly based on intuition" (Adolphs 9). Traditional pragmatics has been discussed in terms of "invented examples of utterances based on native speakers' intuitions" to support a division between form and function (Adolphs 18, 21). These drawbacks lead to a general scepticism of corpora exploration focused on extended discourse stretches (Adolphs).

Nevertheless, Adolphs and O'Keeffe et al. emphasise the importance of spoken corpora for pragmatic investigations. As mentioned above, "much of the work in pragmatics has been based on invented examples of utterances based on native speaker intuition" (Adolphs 21). However, this intuitive aspect can be overcome by the nature of corpora, as they allow us to re-examine the researchers' initial analysis of speech act expressions and to "re-evaluate more traditional frameworks for assigning functions to utterances" (Adolphs 90).

Learner corpora, which are the main focus of this study, contain collections of texts produced by second or foreign language learners. Researchers aim to track the developmental aspects of learners' language and, particularly, to highlight areas which are difficult for learners to learn or acquire (O'Keeffe et al.). Studies of learner corpora predominantly employ frequency-based lexico-grammatical analysis. There is no doubt that interlanguage pragmatics has been heavily based on data collected from elicitation tasks such as the Discourse

Completion Task (DCT) (Kasper and Blum-Kulka; Kasper and Rose; Takahashi; Schauer), while studies based on spoken learner corpora remain scarce.

3. The Aim of the Study: Overcoming Difficulties with Pragmatic Analyses

This paper explains the conversion of TXT files of the NICT JLE Corpus into XML format. As mentioned in Section 2, it is our desire to establish frameworks for interpreting patterns of use that go beyond the lexico-grammar in order to conduct a pragmatic analysis of the NICT JLE Corpus data.

The NICT JLE Corpus may be downloaded for free on the website. It is one of the largest spoken corpora, and is comprised of written transcripts of audio-recorded speech samples of learners at different levels of proficiency. The corpus contains not only information on interactive features such as overlap between speakers, but also data on extra-linguistic features such as fillers, pauses, and repetitions. The marked-up information in the corpus is very useful for investigating pragmatic features. For example, interactive contexts in a role-play session about shopping can be very informative in the identification of requestive speech acts, as they allow the researchers to examine the intended meaning of the speaker.

However, difficulties arise in pragmatic analyses of the NICT JLE Corpus data. First, the corpus analysis tool, called Analyzer, that is provided by the compiler does not easily allow the researchers to amend the data, for example, to separate the data into smaller parts according to the segments annotated or to add more pragmatic annotations. Secondly, only limited frequency counting based on lexical items and tags are available with Analyzer. Finally, even though the TXT files of the corpus data have recently been made freely available online, thereby enabling researchers' direct access, it is still difficult to amend the data. Since the format of annotation tags in the corpus is XML-like but not completely well-formed, it is difficult to use Perl, a programme that allows researchers to amend and segment data in

well-formed XML files.

The remainder of this paper is organised as follows. Section 4 is an introduction to the NICT JLE Corpus, including a description of how and in what format it has been provided by the compiler. After describing the rules of XML files in Section 5, next, in Section 6 we point out examples of ill-formed tags within the corpus and explain the process of converting the ill-formed tags into well-formed XML. Finally, Section 7 discusses how the converted XML may be used for pragmatic analyses, and frequency information on types and tokens is generated.

4. The NICT JLE Corpus

4.1. What is the NICT JLE Corpus?

The NICT JLE Corpus was created by the National Institute of Information and Communication Technology (NICT) in Japan. It contains 1.2 million words in transcripts of 1,281 Japanese EFL learners taking a speaking proficiency test called the Standard Speaking Test (SST) (Izumi, Uchimoto, and Isahara). The SST is a 15-minute oral interview which was developed based on the Oral Proficiency Interview (OPI) of the American Council on the Teaching of Foreign Languages (ACTFL). A total of 300 hours of interviews were transcribed. The SST has five stages: (1) answering warm-up questions (3–4 minutes), (2) describing a single picture (2–3 minutes), (3) doing a role-play with the interviewer (1–4 minutes), (4) narrating picture sequences (2–3 minutes), and (5) answering wind-down questions (1–2 minutes). The participants who took the test were assessed holistically and classified into one of nine proficiency levels called SST Levels: Levels 1 to 3 (Novice), Levels 4 and 5 (Intermediate Low), Levels 6 and 7 (Intermediate Mid), Level 8 (Intermediate High), and Level 9 (Advanced). Stages 2 to 4 consist of the "task" and "follow-up" sessions.

In addition, 167 files of the 1,281 files in the corpus are error-tagged in terms of 47 grammatical and lexical features. Also, for the purpose of comparison with the learner data, a subcorpus containing the data

of 20 native English speakers is available. However, it should be noted that the error-tagged corpus and native corpus are not dealt with in this study, and they are not examined in the conversion to well-formed XML files.

4.2. Tags Annotated in the Corpus

There are 30 basic annotation tags used in the corpus. They can be classified broadly into four types based on the type of information they denote: interview structure, the interviewee's profile, speaker turns, or utterance phenomena such as fillers, repetitions, self-corrections, and overlapping (National Institute of Information and Communication

```
<interview>
<filename></filename>
<head version="1.3">
<date></date>
<sex></sex>
<age></age>
<country></country>
<overseas></overseas>
<category></category>
<step></step>
<TOEIC></TOEIC>
<TOEFL></TOEFL>
<other_tests></other_tests>
<SST_task2></SST_task2>
<SST_task3></SST_task3>
<SST_task4></SST_task4>
<SST_level></SST_level>
</head>
<body basictag_version="2.1.3">
<stage1></stage1>
<stage2>
<task></task>
<followup></followup>
</stage2>
<stage3>
<task></task>
<followup></followup>
</stage3>
<stage4>
<task></task>
<followup></followup>
</stage4>
<stage5></stage5>
</body>
</interview>
```

Fig. 1. Tags for representing the interview structure and the interviewee's profile.

```
<A>How are you?</A>
<B>Fine. Thanks. How are you?</B>
<A>I'm fine, too. Thank you.</A>
```

Fig. 2. Tags for representing speaker turns.

Table 1 Tags for representing utterance phenomena

Tag	Meaning
<F></F>	Filler / Filled Pause
<R></R>	Repetition
<R?></R?>	Repetition (which the transcriber is not confident transcribing)
<SC></SC>	Self-correction
<SC?></SC?>	Self-correction (which the transcriber is not confident transcribing)
<CO></CO>	Utterances which are cut off
<?></?>	Utterances which the transcriber is not confident transcribing
<??></??>	Utterances which are impossible to transcribe
<H pn="X"></H>	Hidden personal information or discriminatory term
<JP></JP>	Japanese
<.></.>	Pause which lasts 2 to 3 seconds
<..></..>	Pause which lasts more than 3 seconds
<OL></OL>	Overlapping utterances of Speaker A and Speaker B
<nvs></nvs>	Non-verbal sounds such as a sniff, laughter, cough, or sigh
<laughter></laughter>	The speaker produces the utterance while laughing.
<ctxt></ctxt>	Non-linguistic events or information to be described

Technology). It should be noted that some of these tags are not considered well-formed XML, as described further in Section 6. Figures 1 and 2 show the first three tag types as listed in NICT's online manual, and Table 1 lists tags of the fourth type.

4.3. How has been the corpus provided?

4.3.1. The NICT JLE Corpus Analysis Tool: Analyzer

Originally, access to the NICT JLE Corpus was only available on a CD-ROM that accompanied the book *Nihonjin 1200 Nin No Eigo Speaking Corpus [L2 Spoken Corpus of 1200 Japanese Learners of English]* (Izumi et al.). In 2012, the TXT files were made available for free online (National Institute of Information and Communication Technology). The CD-ROM contains "The NICT JLE Corpus Analysis Tool" (i.e. Analyzer) with corpus data such as "LearnerOriginal" (i.e. learner data including written interview transcripts of 1,281 interview-

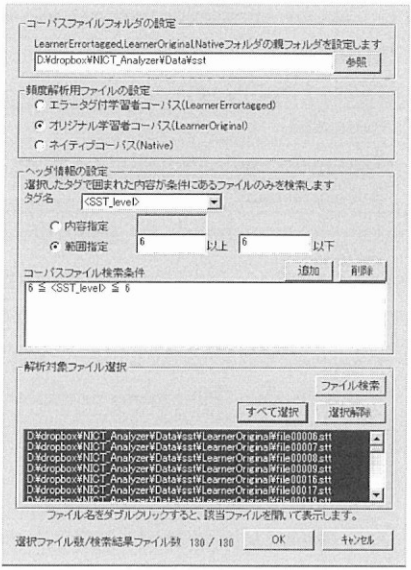

Fig. 3. Analyzer specifying the division of the corpus (Level 6 of LearnerOriginal).

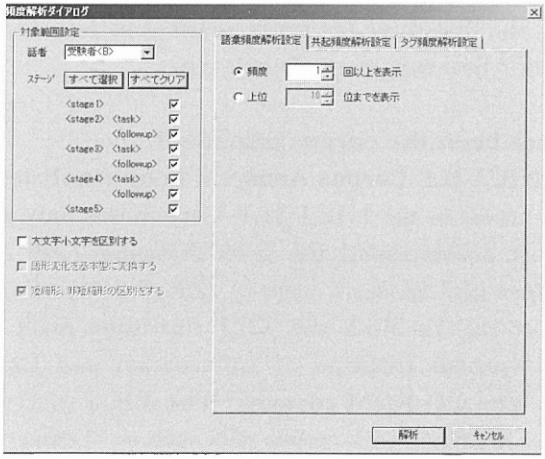

Fig. 4. Analyzer specifying the divisions of the file.

Table 2 The distribution of learners, types, and tokens for each level in the NICT JLE Corpus

SST Level	Proficiency	Participants	Tokens of Speaker A	Tokens of Speaker B
1	Novice Low	3	1,754	413
2	Novice Mid	35	17,980	7,654
3	Novice High	222	103,979	95,494
4	Intermediate Low	482	227,103	308,477
5	Intermediate Low Plus	236	110,603	204,617
6	Intermediate Mid	130	62,563	132,885
7	Intermediate Mid Plus	77	39,872	87,574
8	Intermediate High	56	30,527	70,404
9	Advanced	40	24,204	56,118

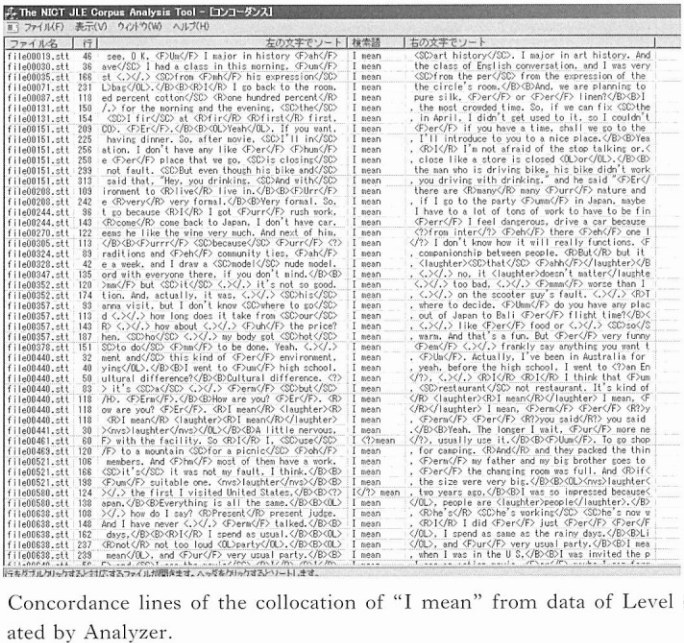

Fig. 5. Concordance lines of the collocation of "I mean" from data of Level 6 generated by Analyzer.

ers and interviewees), "LearnerErrortagged" (i.e. error-tagged learner data), and "Native" (i.e. the data of 20 native speakers taking the SST). These data may only be analysed with Analyzer; they cannot be downloaded as TXT files. Analyzer allows researchers to make a word list of the particular part of the corpus which can be segmented by annotation tags. Figure 3 shows part of the user interface. In this sample window shot, the researcher has specified the corpus version "LearnerOriginal" and a particular part of the corpus (130 files annotated as Level 6), which is segmented by a tag showing the SST Level in the header of each file. Figure 4 shows the interface for generating a word list by specifying the learner (e.g. Interviewee B) and stage (i.e. Stages 1 to 5, including "task" and "follow-up").

Analyzer can easily segment the corpus according to the annotation tags. Then, it generates frequency results and concordance lines depending on the stage or learner proficiency levels, which can also be downloaded as CSV files. Figure 5 shows the concordance lines of "I mean" from the learner data tagged as Level 6. Table 2 shows the distribution of learners (i.e. interviewees), types, and tokens for each level of proficiency based on Analyzer. It should be noted that the types provided are not lemmatised.

However, there are several drawbacks to using Analyzer. First, it is not possible to mark up elements other than tags already annotated in the corpus because the data can only be accessed using this tool. Hence, there is no way to amend the corpus data by annotating additional markup, split into the data into several segments, or even transform the corpus to be analysed by other tools. Next, it is possible to create a word list but not to generate a lemmatised token. Finally, although anyone can download the corpus data from the accompanying CD-ROM and install the software onto his or her own computer, the book is now out of print.

4.3.2. The NICT JLE Corpus Data Provided as Text Files on the Website

In October 2012, NICT began providing the NICT JLE Corpus on

its website (National Institute of Information and Communication Technology). The corpus data are the same as published by Izumi et al. Now, instead of files that can only be processed by Analyzer, the corpus offers each interview transcript in a TXT file. Therefore, researchers can access the data directly and add more tags for further analyses. However, as each file in the corpus contains both an interviewer and interviewee and all the stages of the interview script, it is not possible to retrieve a word list or concordance lines by specifying the data according to speaker tags or stage tags using general concordancers such as AntConc. If researchers wish to focus on specific parts or stages of the interviews in the corpus, the data have to be automatically segmented by other tools. However, there is still a problem with automatic segmenting, as the XML files are not well-formed. Section 5 describes the rules of valid XML files, and Section 6 shows how the corpus data were converted into well-formed XML.

5. The Rules of XML files

XML is defined as "an extensible markup language used for the description of marked-up electronic text" (Sperberg-McQueen and Burnard 13). Markup language means a set of markup conventions used for encoding texts. The purpose of XML is to allow different kinds of processing to "be carried out with the same part of a file" (Sperberg-McQueen and Burnard 14). XML "ensures the documents encoded according to its provisions" (14) and can move "information from place to place, even between different software products and platforms" (Goldfarb and Prescod 6) without loss of information.

Therefore, an XML document can be transported into and processed by any programme without any transformations or translations if it is well-formed. There are three simple rules for writing an XML document. First, there should a single "element" encoded with a start- and end-tag, which is known as the "root element" (Shibano 56). Second, "the tags marking the start and end of each element must always be present" (Goldfarb and Prescod 17). Then, elements should not partially overlap with one another. For example, Shibano writes that

the street name "Rue Slater Street" written in French and English in Canada should be marked up as "<fr>Rue </fr><en><fr>Slater </fr> Street</en>" (59). The markup "<fr>Rue <en>Slater </fr>Street</ en>"is not allowed because the tags overlap; that is, the start-tag <en> is opened within <fr></fr> but closed outside it. The uppermost elements (i.e. parent node) should always contain the lower elements (i.e. child nodes).

Once an XML document is confirmed to be well-formed, it is called a "valid document", and a document which states the criteria for successful validation is known as document type declaration (DTD) or an XML schema (Sperberg-McQueen and Burnard 17–18; see also Shibano 66–67; Goldfarb and Prescod 15–16, 40, 42). An element can have attribute values. "Attribute-value pairs" can be found inside the start-tag "<poem id='P1' status="draft">", where "the value part must always be given inside matching quotation marks, either single or double" (Sperberg-McQueen and Burnard 22). On the other hand, an end-tag does not contain an attribute value specification, as illustrated by "</poem>" (22).

According to various websites that list XML rules, such as W3C (Bray, Hollander, and Layman), there are also several rules regarding elements:

i. The first character should be "_" (underscore), ":" (colon), or one-bite English letter or Japanese letter (except for one-byte kana characters and two-byte English letters or numbers).
ii. Characters such as a one-byte number, "." (full stop), "-" (hyphen), and letters with accent symbols should be used from the second letter.
iii. One-byte kana characters, two-byte English letters or numbers, and two-byte spaces cannot be retrieved.
iv. Reserved keywords such as "xml" cannot be used.

6. Data Cleansing with the NICT JLE Corpus

To convert the NICT JLE Corpus into well-formed XML format, we performed automatic modification in Perl and manual modification

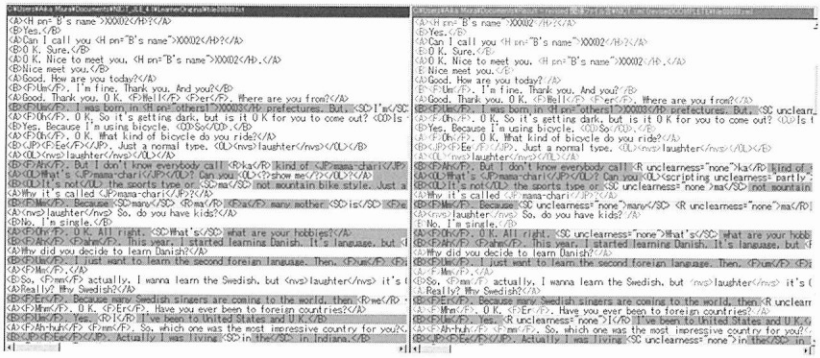

Fig. 6. The original and modified versions of "file00008" shown in WinMerge.

by opening the files in Chrome, as "Web browsers will display ill-formed documents" (Goldfarb and Prescod 42). Thus, modification of overlapping tags, modification of missing or erroneous tags, and correction of the order of end-tags were performed manually. Figure 6 shows an example of these changes; the original file is on the left, and the modified version on the right. Darker shaded-areas on the window illustrate the modified parts, including tags highlighted with a lighter shade.

6.1. Automatic Modification of Ill-formed Tags

Table 3 lists the tags that were converted into XML tags with elements and attributes because symbols such as "." (full stop) and "?" (question mark) cannot be used. The conversion was automatically processed using Perl. Eight tag sets were converted.

6.2. Manual Modification with Ill-formed Tags

6.2.1. Modification of Overlapping Tags

As Shibano states, overlapping tags are not allowed. However, there are two types of overlapping tags in the NICT JLE Corpus: tags which overlap utterances by different speakers and those which intersect with other tags within the same speaker's utterance. After the start-tags of <R> (repetition) and <SC> (self-correction) were auto-

Table 3 Tags automatically converted into well-formed XML

Original Tag	Modified Tag	Notes
<.></.>	<pause duration="long"></pause>	The tags showing pauses, <.> and <..>, are combined. To distinguish the length of pause, attribute values are added.
<..></..>	<pause duration="short"></pause>	
<?></?>	<scripting unclearness="partly"> </scripting>	The tags showing how confident the transcriber is in the transcription, <?> and <??>, are combined. To distinguish the degree of confidence, attribute values are added.
<??></??>	<scripting unclearness="all"> </scripting>	
<SC></SC>	<SC unclearness="none"></SC>	The tags showing self-correction, <SC> and <SC?>, are combined. To distinguish the degree of the transcriber's confidence in the transcription, attribute values are added.
<SC?></SC?>	<SC unclearness="partly"></SC>	
<R></R>	<R unclearness="none"></R>	The tags showing repetition, <R> and <R?>, are combined. To distinguish the degree of the transcriber's confidence in the transcription, attribute values are added.
<R?></R?>	<R unclearness="partly"></R>	

Table 4 Overlapping tags (between speakers) manually converted into well-formed XML

Original Tag	Modified Tag	Number of Modified Tags
<CO></CO>	<CO segment="inter">	48
<R unclearness="none">	<R unclearness= "none" segment="inter">	8
<SC unclearness="none">	<SC unclearness= "none" segment="inter">	50

matically modified to include attributes, the tags below were manually modified and checked for validity in Chrome. For example, the original version of "file00074" had an overlapping tag "repetition". Chrome indicated that there was an "error on line 189 at column 110: Opening and ending tag mismatch: R line 0 and B". Figure 6.1 shows the original data, in which the start-tag <R> in line 189 does not have the end-tag </R> before the end-tag of speaker </B>. Instead, the end-tag

```
189 <B><F>Mhm</F> I watched "Life Is Beautiful". <F>Mhm</F>. And on Sunday, <R><OL>I went
    to</OL></B>
190 <A><OL><F>Uh-huh</F></OL>.</A>
191 <B>Yokohama</R> I went to Yokohama.</B>
```

Fig. 6.1. The original "file00074" showing the overlapping tags between speakers.

```
189 <B><F>Mhm</F> I watched "Life Is Beautiful". <F>Mhm</F>. And on Sunday, <R unclearness="none"
    segment="inter"><OL>I went to</OL></R></B>
190 <A><OL><F>Uh-huh</F></OL>.</A>
191 <B><R unclearness="none" segment="inter">Yokohama</R> I went to Yokohama.</B>
```

Fig. 6.2. The modified version of "file00074".

Table 5 Overlapping tags (single speaker's utterance) manually converted into well-formed XML

Original Tag	Modified Tag	Number of Modified Tags
<CO>	<CO segment="intra">	1
<OL>	<OL segment="intra">	3
<SC>	<SC unclearness= "none" segment="intra">	8
<R>	<R unclearness= "none" segment="intra">	9

appears inside the next speaker tags but without the start-tag in line 191. Therefore, the end-tag in the first utterance of Speaker B in line 189 and start-tag in the second utterance in line 191 were added as in figure 6.2.

There are also tags which overlap other tags within the same speaker's utterance, as shown in table 5. For example, as shown in figure 7.1, the original version of "file00287" is not valid, as Chrome showed an "error on line 40 at column 33: Opening and ending tag mismatch: SC line 0 and OL". The tags "overlap with other speakers" and "self-correction" overlap in the utterance of Speaker B. Figure 7.2 shows that not only was the start-tag "self-correction" modified with additional attributes in line 40, but also the start- and end-tags were inserted after </OL>.

```
40 <B><OL><SC>it's</OL> kind of</SC> <F>well</F> it's nice and safe. <nvs>laughter</nvs></B>
41 <A><F>Uhmm</F>. <OL><CO>I</CO></OL>.</A>
```

Fig. 7.1. The original version of "file00287" showing overlapping tags within Speaker B's utterance.

```
40 <B><OL><SC unclearness="none" segment="int ra">it's</SC></OL> <SC unclearness="none"
   segment="intra">kind of</SC> <F>well</F> it's nice and safe. <nvs>laughter</nvs></B>
41 <A><F>Uhmm</F>. <OL><CO>I</CO></OL>.</A>
```

Fig. 7.2. The modified version of "file00287".

```
And <.></.> <R>two</R> two wòman are talking. <F>Mm</F>. And <.></.> <F>uhm</F>
<laughter><R>bo</R> <SC>boys</laughter> <F>uhmm</F> playing</SC> boys are playing volleyball.
```

Fig. 8.1. The original line 97 from "file00141" showing overlapping tags within the same utterance.

```
And <pause duration="short"></pause> <R unclearness="none">two</R> two woman are talking. <F>Mm</F>.
And <pause duration="short"></pause> <F>uhm</F> <laughter><R unclearness="none">bo</R></laughter> <SC
unclearness="none"><laughter>boys</laughter> <F>uhmm</F> playing</SC> boys are playing volleyball.
```

Fig. 8.2. The modified version of line 97 from "file00141".

Next, the original corpus contains 16 "laughter" tags that overlap either between different speakers' utterances or within the same utterance. The modification of "file00141" illustrates how this problem was solved. Chrome displayed the message "error on line 97 at column 509: Opening and ending tag mismatch: SC line 0 and laughter" in a part "<SC>boys</laughter> <F>uhmm</F> playing</SC>", as shown in figure 8.1. The end-tag </laughter> after "bo</R>" and start-tag <laughter> before "boys" were added, as shown in figure 8.2. While "self-correction" is supposed to mark certain lexical items (in this case, "boys uhmmm playing"), "laughter" is an additional non-linguistic annotation that occurs simultaneously with the utterances of lexical items. In the process of modification, only the first end-tag and the second start-tag of "laughter" were inserted to resolve the problem of overlapping tags.

6.2.2. Provision of Missing Tags

There were four tags missing in the original corpus that were added in during the modification process. They are shown in table 6.

6.2.3. Correction of Erroneous Symbols

Erroneous symbols were found in the original corpus, as table 7 shows. First, two-byte characters or spaces were corrected to one-byte ones. Second, irrelevant symbols were deleted.

6.2.4. Modification of the Wrong Order of End-Tags

Eighty-five parts in the original corpus which had the wrong order of end-tags, as in the example of line 60 from "file00008" in figure 9.1.

Table 6 Modification of missing tags

Missing start-tag	<A>	file00062	line 37
	<B>	file00003	line 47
	<B>	file00794	line 112
	<B>	file00963	line 183

Table 7 Modification of erroneous tags

Two-byte characters	I	file00081	line 69
	space	file01165	line 157
	space	file01165	line 157
Irrelevant symbols	.	file00743	line 14
	.	file01270	line 14
	]	file00848	line 36

```
<SC>very</SC> totally different from Japan, so I really like that place even though it's <SC>a
<laughter>very</SC></laughter>
```

Fig. 9.1. The original line 60 from "file00008" showing the wrong order of end-tags.

```
<SC>very</SC> totally different from Japan, so I really like that place even though it's <SC unclearness="none">a
<laughter>very</laughter></SC>
```

Fig. 9.2. The modified version of line 60 from "file00008".

Figure 9.2 shows the corrected order of the end-tags "laughter" and "self-correction".

7. Application of Well-Formed XML Files in the Analysis of Longer Stretches of Discourse in the NICT JLE Corpus

7.1. Research on Request Strategies in Speech Acts

As described in Section 3, researchers now have direct access to the data and markup annotations of the NICT JLE Corpus, thanks to NICT providing TXT files. Pragmatic annotations were manually added to the segmented parts of the valid XML files in the corpus, following the established rules of XML format (Miura, *Criterial Features; The NICT JLE*). Her studies attempt to investigate request strategies in speech acts in the extracted learner data of role-play sessions that involve transactions to obtain goods, negotiations for refunds, or item exchanges. In this role-play (segmented by the tag <stage3> and </stage3>), the interviewer plays a shop assistant or train staff member, while the interviewee is given the role of a customer or passenger. Figure 10 shows an example from "file00001" with added annotations of request strategies (shown in bold). The annotation was done according to the coding scheme developed in the area of cross-linguistic pragmatics (Blum-Kulka; Blum-Kulka, House, and Kasper; Trosborg; Salgado).

As figure 10 shows, the *head act* (i.e. the core of the request sequence) of request strategies was first identified and tagged as <HA></HA>.

```
121 <A>Hello. May I help you, miss?</A>
122 <B><F>Er</F> yes. <F>Mmm</F> <HA><RQ dmc="s"><DR str="desire"><R unclearness="none">I
    want to</R> <SC>I want to</SC> <SC>I want</SC> <F>mm</F> sorry, <F>mm</F> <pause
    duration="short"></pause> I want to <F>err</F> watch.</DR></RQ></HA></B>
123 <A><F>Uhm</F>.</A>
124 <B><F>Um</F>.</B>
125 <A>Yes.</A>
126 <B><F>Um</F> and <F>mmm</F> <HA><RQ dmc="s"><ID>I prefer <F>mm</F> leather
    watch.</ID></RQ></HA></B>
127 <A><F>Uhm</F>.</A>
128 <B><F>Uhm</F> and <F>mmm</F> <HA><RQ dmc="h"><ID><SD mkr="intrg"><R
    unclearness="none">do you have</R> <F>mm</F> do you have something special
    one?</SD></ID></RQ></HA></B>
```

Fig. 10. An excerpt from "file00001" with annotations of request strategies.

Then, whether the dominance of the requestive perspective is on the speaker or hearer was identified (<RQ dmc="s"> or <RQ dmc="h">). There are three types of request strategies: *direct, conventionally indirect* and *non-conventionally indirect* strategies. In this example, the direct request strategy is expressed with a pattern of desires ("I want/need X") and annotated as <DR str="desire"> in line 122. Then, head acts of request in lines 126 and 128 were identified as an indirect strategy, annotated as <ID>. The syntactic downgrader as interrogative as internal modification was also identified as <SD mkr="intrg">.

After manual tagging of these pragmatic features, annotated features were retrieved using Perl. The Perl script was written to retrieve not only the search tags, but also the whole utterance of a particular line where the search tags appeared, as well as the line number and file identification number. This made easier to examine the neighbouring contexts of the target pragmatic features, as we could retrieve longer stretches instead of a KWIC (Key Word In Context) concordance with a limited number of words provided by general concordancers. It was also possible to check the preceding and following utterances of the learners and their interactions with the interviewers, by detecting the place where the target features were produced according to the given line numbers.

In summary, the valid XML documents of the NICT JLE Corpus allow us to expand the scope of analysis of pragmatic features in longer stretches of discourse, rather than restricting the scope to word frequency or simple concordance lines; this means lexical and grammatical analyses are no longer limited to surface forms, as when Analyser or general concordancers are used. The details of coding scheme and results of analyses drawing on different proficiency levels and tasks given role play sessions can be found in Miura's studies (*Criterial Features; The NICT JLE*). It is also reported that the development and availability of spoken learner corpora not only give new insights into the development of tools or coding schemes for analysing the relationship between lexico-grammatical features and discourse functions, but also allow us to re-examine the results of the previous stud-

ies on interlanguage pragmatics.

7.2. Word List Based on Well-Formed XML Files in Comparison to Analyzer

Now the valid XML documents are available for the NICT JLE Corpus, the originally designed programme can retrieve its tokens and lemmatised types of learner data at different proficiency levels. The following programme was made to generate a word list and frequency information.

i. Elements marked up by speaker tags "A" and "B" were extracted respectively.
ii. Elements annotated by tags such as "F", "H", "ctxt", and "nvs" were excluded and those annotated by "laughter", "R", "CO", and "OL" were counted.
iii. Elements segmented by a space, comma (,) and question mark (?), and exclamation mark (!), double quotation (" "), and semicolon (;) were identified as lexical words. In this case, upper case and lower case letters were not distinguished.
iv. Tokens and types for proficiency level and speaker were counted. E_lemma.txt (Ver1.1), which was compiled by Yasumasa Someya (Izumi et al.), was used to lemmatize the retrieved types. Inflected verbs and pluralised nouns were converted into dictionary forms in the list, but the part-of-speech information were not considered.
v. Genitive forms, contracted forms, or hyphenated words were not deconstructed, and counted as single lexical words.

Table 8 shows a comparison between the distribution of learners, types, and tokens for each level of proficiency retrieved from Analyzer and XML files using the Perl programme. It should be noted that Analyzer cannot generate lemmatised types. The differences in the tokens and types of the present study from those of Analyzer may be due to the way of counting of contracted forms, genitive forms, hyphenated words, and words annotated as hidden information, and the process of lemmatisation. We aim to improve the precision of the

Table 8 The Distributions of Learners, Types, and Typees for Each Level in the NICT JLE Corpus

SST Level	Analyzer's Tokens	Analyzer's Types Types (not lemmatised)	Tokens in the Present Study	Lemmatised Types in the Present Study
1	413	208	411	238
2	7,654	1,259	7,498	1,408
3	95,494	4,670	94,290	4,946
4	308,477	7,410	306,243	7,540
5	204,617	5,893	203,146	5,669
6	132,885	5,034	130,287	4,611
7	87,574	3,953	85,018	3,582
8	70,404	3,607	68,349	3,220
9	56,118	3,264	54,251	2,840

programme for generating a word list and frequency in the future.

8. Conclusion

The automatically and manually converted 1,281 files into well-formed XML format have made it possible to amend the corpus data with any tools in a way that researchers wish. Especially, pragmatic analyses can be more easily done with valid XML versions. In order to investigate instances of requestive speech acts in the role-play stage of learner data, it was highly useful to retrieve a word list or frequency information from the segmented corpus data according to the annotation tags, for example, stages of the interview, proficiency levels of learners, turns of interviewers and interviewees (i.e. learners). Interactive spoken data with abundantly annotated extra-linguistic elements of the XML-formatted version of the NICT JLE Corpus allow researchers to expand their analysis from surface forms to pragmatic functions in longer stretches of discourse.

WORKS CITED

Adolphs, Svenja. *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: John Benjamins, 2008. Print.

Blum-Kulka, Shoshana. "Indirectness and Politeness in Requests: Same or Different?" *Journal of Pragmatics* 11 (1987): 131–46. Print.

Blum-Kulka, Shoshana, Juliane House, and Gabriele Kasper. *Cross-cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex, 1989. Print.

Bray, Tim, Dave Hollander, and Andrew Layman, eds. "Namespaces in XML." *XML Schema*. W3C, 14 January 1999. Web. 8 July 2014. <http://www.w3.org/TR/1999/REC-xml-names-19990114>.

Goldfarb, Charles F., and Paul Prescod. *The XML Handbook*. Upper Saddle River, NJ: Prentice Hall, 2011. Print.

Izumi, Emiko, Kiyotaka Uchimoto, and Hitoshi Isahara, eds. *Nihonjin 1200 Nin No Eigo Speaking Corpus. [L2 Spoken Corpus of 1200 Japanese Learners of English]*. Tokyo: ALC, 2004. Print.

Kasper, Gabriele, and Shoshana Blum-Kulka. "Interlanguage Pragmatics: An Introduction." *Interlanguage Pragmatics*. Ed. Gabriele Kasper and Shoshana Blum-Kulka. Oxford: Oxford University Press, 1993, 3–17. Print.

Kasper, Gabriele, and Kenneth R. Rose. *Pragmatic Development in a Second Language*. Oxford: Blackwell, 2002. Print.

Knight, Dawn, and Svenja Adolphs. "Multi-modal Corpus Pragmatics: The Case of Active Listenership." *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Ed. Jesús Romero-Trillo. Berlin: Mouton de Gruyter, 2008. 175–90. Print.

Miura, Aika. "Criterial Features of Pragmatic Competence in a Spoken Corpus of Japanese Learners of English to Profile Different Levels of Proficiency." Learner Corpus Research Conference (LCR 2013). Solstrand Hotel & Bad, Bergen. 27 Sep. 2013. Address.

Miura, Aika. "The NICT JLE (Japanese Learner English) Corpus No XML Seikeishikika To Sore Wo Tsukatta Shutokubetsu No Goyoronteki Gengo Tokuchou No Bunseki [Constructing Well-Formed XML Files of the NICT JLE Corpus and Its Application to Research into Pragmatic Competence Across Different Levels of Proficiency]." The 39th Conference of the Japanese Association for English Corpus Studies. Tohoku University, Sendai. 5 Oct. 2013. Address.

National Institute of Information and Communication Technology. *The NICT JLE (Japanese Learner English) Corpus*. Information Analysis Laboratory. Web. 15 Feb. 2014. <http://alaginrc.nict.go.jp/nict_jle/index_E.html>.

O'Keeffe, Anne, Brian Clancy, and Svenja Adolphs. *Introducing Pragmatics in Use*. Abingdon: Routledge, 2011. Print.

Romero-Trillo, Jesús. "Introduction: Pragmatics and Corpus Linguistics—a Mutualistic Entente." *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Ed. Jesús Romero-Trillo. Berlin: Mouton de Gruyter, 2008. 1–10. Print.

Salgado, Elizabeth Flores. *The Pragmatics of Requests and Apologies: Developmental Patterns of Mexican Students*. Amsterdam: John Benjamin Publishing Company, 2011. Print.

Schauer, Gila A. "Pragmatic Awareness in ESL and EFL Contexts: Contrast and Development." *Language Learning* 56 (2006): 269–318. Print.

Shibano, Koji. *SGML/XML Ga Wakaru Hon [SGML/XML]*. Tokyo: Ohmsha, 2000.

Print.

Sperberg-McQueen, C. M., and Lou Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange: TEI P4*. Oxford: Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford, 2002, Print.

Takahashi, Satomi. "Pragmatic Transferability." *Studies in Second Language Acquisition* 18 (1996): 189–223. Print.

Trosborg, Anna. *Interlanguage Pragmatics: Requests, Complaints and Apologies*. Berlin, NY: Mouton de Gruyter, 1995.