Design of the ICNALE-Spoken : A New Database for Multi-modal Contrastive Interlanguage Analysis

Ishikawa, Shin'ichiro

(Citation)
Learner Corpus Studies in Asia and the World,2:63-75

(Issue Date)
2014-05-31

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81006690

(URL)
https://hdl.handle.net/20.500.14094/81006690

Design of the ICNALE-Spoken: A New Database for Multi-modal Contrastive Interlanguage Analysis

Shin'ichiro ISHIKAWA
Kobe University

Abstract

The current paper introduces the principle, design, and development procedure of the ICNALE-Spoken, a new learner speech corpus. The ICNALE-Spoken contains highly controlled spoken interlanguage productions by varied Asian learners as well as the comparable spoken data by English native speakers. Using the ICNALE-Spoken in combination with the ICNALE-Written (Ishikawa, 2013), which was developed under the same principle, makes it possible for us to conduct a reliable spoken-written multi-modal contrastive interlanguage analysis. In order to outline the unique features of the ICNALE-Spoken, we will compare it with The LINDSEI as an archetypal learner speech corpus with paying a special attention to target learners, task design, data size, and data distribution.

Keywords

Learner corpus, Speech data collection, Multi-modal contrastive interlanguage analysis

I Introduction

The author released the International Corpus Network of Asian Learners of English (ICNALE) in 2013. The ICNALE is a collection of 1.3 million words of argumentative essays written by college students in ten countries and areas in Asia as well as by English native speakers. The ICNALE is one of the largest learner corpora publicly available and practically the sole database focusing on Asian learners' interlanguage use.

The ICNALE aims to be a reliable database for varied contrastive analyses. Although we can point out many differences by comparing large quantities of learner essays, it is not necessarily easy to identify what causes the observed differences because learners' interlanguage use is highly prone to exhibit varied extra-textual elements. Therefore, in comparison to other major learner corpora such as the International Corpus of Learner

English (ICLE) (Granger *et al.* 2002/ 2009), the ICNALE contains more highly controlled interlanguage productions. For example, the topics, the time, the length, and the reference use are all controlled in a rigid way. This homogeneity of the collected data guarantees reliability in contrastive analyses of learners' interlanguage outputs.

The wide-spread use of the ICNALE in branches of applied linguistics, however, has shown its essential drawback, namely the lack of spoken data. Without analysing both learners' spoken and written productions, we cannot discuss learners' interlanguage use in general.

Thus, we have launched a new project to collect the spoken productions by Asian learners of English to compile the ICNALE-Spoken. Comparison of the ICNALE-Spoken with the ICNALE-Written (the former "ICNALE") makes it possible to conduct a reliable spoken-written multi-modal contrastive interlanguage analysis, which sheds new light on studies of interlanguage use by international learners.

II An Archetypal Learner Speech Corpus: The LINDSEI

Learner speech corpora are more limited in number and variety in comparison to their written counterparts. In fact, there is only one significant speech corpus which includes rich international data and which is publicly available: the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Guilquin *et al.* 2010).

Although the LINDSEI emphasizes naturalness of data rather heavily as illustrated below and although it is clearly a different type of corpus from the ICNALE, which emphasizes control of data, the design and structure of the LINDSEI as a pioneer work offers many hints for corpus developers.

2.1 Target Learners

The LINDSEI collects the spoken data of 554 advanced-level college students with Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish L1 backgrounds. As ten of these L1 groups (all except Greek speakers) are also covered in the ICLE, we can compare spoken and written productions by those learners. The number of participants for each L1 group is approximately 50.

When analysing learners' interlanguage production, it is crucial to pay due attention to their backgrounds. Thus, the LINDSEI collects learners' profile data, such as their learning context, proficiency, age, gender, L1, country, other foreign languages, and length of stay in English-speaking countries.

2.2 Task Design

The LINDSEI collects spoken data produced in a kind of oral proficiency interview, which comprises three stages: a warming-up (set topic), a free informal discussion, and a picture description. First, learners choose one of three set topics: "an experience you

have had which has taught you an important lesson," "a country you have visited which has impressed you," and "a film/play you have seen which you thought was particularly good/bad"; after some preparation, they talk about the chosen topic for three to five minutes. Next, they are asked to answer several questions posed by an interviewer. Some of the questions concern the chosen topic, but others are about more general issues like college life, a hobby, an experience of traveling abroad, and so on. Last, they are presented four pictures and asked to describe them. All the speech data included in the LINDSEI are thus part of a dialogue, not a monologue.

The interview lasts for about fifteen minutes, but neither the total time nor the time for each stage is rigidly controlled. For instance, the proportion of speeches discussing "a country you have visited" is 32% for an L1 Chinese group and reaches 68% for an L1 German group in terms of the number of words.

The goal of the LINDSEI is to collect speech data produced in an "informal" setting and with "few constraints on language production." The LINDSEI team notes that "[o]nly the open-ended elicited data such as written compositions or oral interviews qualify as learner corpus data" (Guilquin *et al.* 2010, p.5); and they also add that the speeches collected in the warm-up (set topic) and free discussion stages are more "natural" in that "learners were free to choose both the ideas they wanted to express and the language in which to express them" (*ibid.*, p.6). From their viewpoint, both the LINDSEI and the ICLE collect the learner data produced with the least constraints, which they believe guarantees comparability of the two corpora.

> One of the key assets of LINDSEI is that it can easily be compared with its sister corpus, ICLE, which has been compiled on the basis of similar design criteria and partly covers the same mother tongue backgrounds.... For the first time, it will be possible to compare the written and spoken productions of higher intermediate to advanced learners of English on the basis of large electronic collections of language use data. (Gilquin *et al.* 2010, p.4)

2.3 Data Size

The corpus contains 1,079, 681 words of data in total, but the amount of utterance by learners ("B turns only") is 792, 141 words. The average amount of utterance by a single speaker is 1,430 words, and that of a single L1 group is 72,000 words (calculated from the data given in Guilquin *et al.* 2010, p.25).

However, as easily expected, the amount of utterance greatly varies according to L1 groups. For instance, a Japanese learner produces only 728 words, while a Polish learner produces 1,862 words in the same duration of the interview.

2.4 Data Distribution

As in many L1 speech corpora, the LINDSEI distributes only the transcripts, not the

sound data itself. So that corpus users can better understand the properties of the actual speech even without the original sound file, the LINDSEI team has devised a new speech transcription standard with reference to a simple and readable framework proposed in Edwards (1992).

The below is a sample of the transcripts collected in the corpus:

(A(have you decided your topic (?A(
(B(yes I decided topic one mm I'll talk about when when I was high school student second year I went to Australian for one and half month (?B(
(A(mhm (?A(
(B(it's er I stayed Australian family's house and I went to er Australian high school. I joined first grade class mm mm I I joined same class all same class (?B(
(A(uhu (?A(
(B(of them (?B((LINDSEI-JP=027)

Fig. 1 Sample of the data in the LINDSEI

The transcription standard of the LINDSEI defines how to express empty pauses (*e.g.* "..."), filled pauses and backchannelling (*e.g.* "(eh)"), unclear passages (<X>), uncertain words or word endings (<?>), truncated words (*e.g.* "Ger=" and "Germany"), contracted forms (*e.g.* "I'm"), non-standard forms (*e.g.* "cos"), foreign words (*e.g.* <foreign> Traducciión e Interpretaciión </ foreign>), abbreviations (*e.g.* "NASA"), dates and numbers (*e.g.* "nineteen eighty-one"), lengthened syllables (*e.g.* "went to:"), pronunciation of articles (*e.g.* "a [ei] parody," "the [i:] the sketch"), foreign pronunciation (*e.g.* <foreign> distinction </ foreign>), anonymisation (*e.g.* "Hi, <first name of interviewee>"), speaker/ task delimitation, speaker turns (<A>, <B>), overlapping (<overlap>), voice quality (e.g. <starts laughing>, <starts whispering>), non-verbal sounds (*e.g.* <coughs>), and contextual comments (*e.g.* <someone enters the room>).

III Development of the ICNALE-Spoken

Any developer of a learner corpus is often faced with the choice between naturalness and control in data collection. The former contributes to a better representation of the actual language use, while the latter contributes to greater reliability in a contrastive study. They are often in a trade-off relation.

In contrast to the LINDSEI and other learner corpora, the ICNALE, including both of the written and spoken modules, explicitly adopts the latter principle and attempts to collect the data as homogeneously as possible.

3.1 Target Learners

The ICNALE-Spoken plans to collect speech data from leaners in ten countries and

areas in Asia—China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore, Taiwan, and Thailand—as well as from English native speakers. The coverage of learners is the same as that in the ICNALE-Written. All three of the so-called concentric circles of English speaker (Kachru, 1992), namely, the Inner, Outer, and Expanding Circles, are included in both of the corpora. This makes it possible for us to conduct not only a spoken–written multi-modal contrastive interlanguage analysis for Asian learners but also an analysis of the varieties of world Englishes in the region. The number of participants for each learner group is 50 or 100.

The data of 250 learners and 50 English native speakers is to be released for evaluation as the ICNALE-Spoken Baby (Version 1) by the summer of 2014. As each subject produces four speeches (See 3.2), the total number of speeches included in the Baby corpus amounts to 1,200. Learner data includes the speeches by 50 subjects from China, Japan, Taiwan, and the Philippines respectively; and native speaker data includes the speeches by 8 subjects from Australia, 6 subjects from Canada, 10 subjects from UK, 1 subject from New Zealand, and 25 subjects from USA.

Following the subjects' profile investigation scheme developed for the ICNALE-Written, the ICNALE-Spoken also contains very detailed background information on the learners, which is gathered with the ICNALE Learner Profile Questionnaire distributed in MS Excel format.

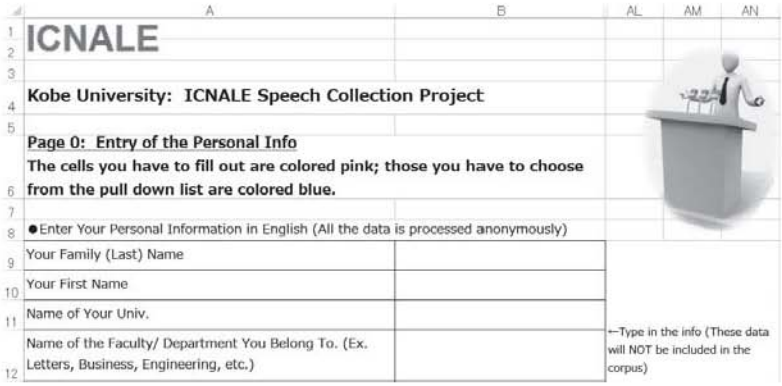

Fig. 2 The ICNALE Learner Profile Questionnaire

The collected profile data include learners' sex, age, college grade (*e.g.* freshman, sophomore), number of years of English study, stay in English-speaking countries, college major, scores on standard English proficiency tests (*e.g.* TOEIC, TOEFL) and/or standard vocabulary size tests, proficiency level based on the Common European Framework of Reference (CEFL) estimated from the test scores, type of motivation for learning English (namely, integrative vs. instrumental), English learning style at

primary, secondary, and college levels, English use outside of school, type of skill that was the focus of study (*e.g.* reading, speaking), amount of instruction by a native speaker, and specialized instruction in pronunciation, presentation, and essay writing (Ishikawa, 2013).

The ratios in proficiency bands and sexes in different modules of the the ICNALE-Spoken Baby are shown below. The four proficiency bands, A2 (Waystage), B1_1 (Threshold lower), B1_2 (Threshold upper), and B2+ (Vantage or Higher), are roughly equal to 225+, 550+, 670+, and 785+ in TOEIC Test; 57+, 72+, 87+, and 110+ in TOEFL iBT Test.

Table 1 Ratios of Proficiency Bands and Sexes in the ICNALE-Spoken Baby (Number of subjects and speech samples)

Modules	Proficiency Bands				Sexes	
	A2	B1_1	B1_2	B2+	Male	Female
China	2 (8)	14 (56)	29 (116)	5 (20)	38 (152)	12 (48)
Japan	10 (40)	20 (80)	10 (40)	10 (40)	31 (124)	19 (76)
The Philippines	0 (0)	6 (24)	39 (156)	5 (20)	21 (84)	29 (116)
Taiwan	9 (36)	21 (84)	9 (36)	11 (44)	8 (32)	42 (168)
Native Speakers	--	--	--	--	43 (172)	7 (28)

It is expected that the data becomes more balanced as the collected data increases in size. By using these detailed learner profiles, corpus users can select a group of highly similar learners across different countries to conduct reliable comparative studies.

3.2 Task Design

Task design should be determined by the type of data to be collected. Unlike the LINDSEI, which collects learners' dialogues, the ICNALE-Spoken collects their monologues. This is because we take the view that monologues, which are not facilitated by an interviewer, are qualitatively more similar to the essays collected in the written corpus, which are produced solely by the learner.

The topics adopted for the ICNALE-Spoken are the same as those used in the ICNALE-Written. Learners are given two topics: "It is important for college students to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country," and are then asked to state whether they agree or disagree with the statements by providing reasons and specific details to support their own claims."

In order to determine the length of a single speech, we conducted several pre-experiments with Japanese learners at varied proficiency levels and found out that even learners at the B2 level or higher often cannot continue to speak for more than 1 minute. Thus, the length of a speech was finally set as 60 seconds. All the learners are asked to continue to speak until the time is up. When learners end their speeches earlier, their

data are all excluded from the corpus.

Another finding obtained from the pre-experiments is that learners, even if they are given enough time for preparation, cannot speak well and much enough in the first trial, but many of them perform more and much better in the repeated trials. Thus, we have decided to give all the learners two chances to talk about the same topic. Repeating the same points is not prohibited, but learners generally use somewhat different lexical items and phrases even when trying to convey a similar idea.

Based on a series of pre-experiments, the task design of the ICNALE-Spoken has finally been fixed as follows:

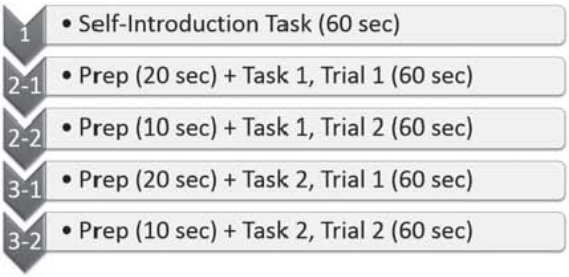

Fig. 3 Task design of the ICNALE-Spoken

First, learners are given a self-introduction task and asked to talk about themselves freely. This task is intended to make the subjects feel relaxed and get a feel for the length of 60 seconds before they conduct a topic speech; the data from the task is not included in the corpus. Next, learners are given the first topic orally ("a part time job for college students") and asked to talk about it after preparation, and then the same task is repeated. Finally, they are given the second topic ("non-smoking in restaurants") and the task is repeated. After finishing all the speeches, learners are told to self-evaluate their own speeches on the scale 0 (Very bad) to 3 (Very good). See the Appendix for the complete structure of the data collection protocol.

Another thing to be considered when collecting data for the compilation of a learner corpus is how to collect data in different countries and areas and often in different periods of time. In case of the LINDSEI, local organizers engaged in the project are responsible for conducting the interviews as well as collecting data. However, this approach is not adopted in the ICNALE-Spoken. This is because it is practically impossible to get the interviewers to conduct all the interviews in the exact same way at any time or place.

Instead, we have developed the ICNALE-Spoken Automatic Speech Collection System (ASCS), which integrates a toll-free international calling line, an answering phone system, and a cloud data storage system.

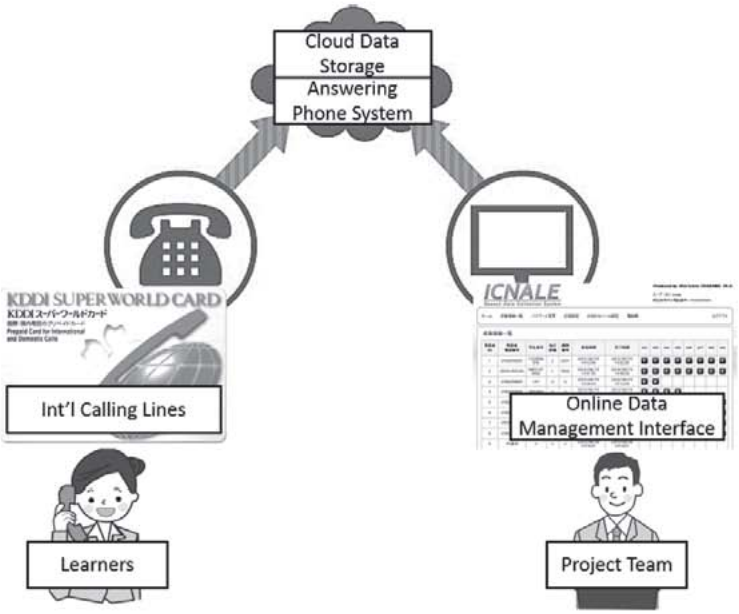

Fig. 4 Outline of the ICNALE-Spoken Automatic Speech Collection System (ASCS)

When learners make a phone call to the designated number, they first hear several messages prompting them to say their name, college, and student number. Next, they hear an instruction to begin speaking about the first topic at the sound of the beep. After the pre-set preparation time, the beep sounds and they immediately begin speaking. After 60 seconds, recording automatically stops. Then, they hear another instruction to speak about the same topic again. Repeating the same or similar content is not prohibited. After the reduced preparation time, they begin to talk about it. This two-trial process is then repeated for the second topic.

Learners' speeches are automatically recorded and individually stored as different wav files. The project team can access all data through the online data management interface and download complete speeches or individual parts anytime (Fig. 5).

The ASCS, which enables us to give the same instructions in the same timeframe to all subjects in spite of different locations and times, guarantees the homogeneity of the data collected in the ICNALE-Spoken.

Fig. 5 Online data management interface of ASCS

3.3 Data Size

We have already collected more than 100,000 words of learner speech from several countries and areas in Asia. The amount of speech included in the ICNALE-Spoken Baby is shown in Table 2.

Table 2 Size of the Modules in the ICNALE-Spoken Baby

Modules	Tokens	Types
China	20,304	1,289
Japan	14,287	847
The Philippines	25,169	1,752
Taiwan	19,430	1,177
Native Speakers	30,867	1,992
Total	110,057	3,384

As the table shows, the size of the data collected in the ICNALE-Spoken is generally smaller than that in the LINDSEI. This is mainly because the ICNALE-Spoken collects controlled monologues by learners. The table also suggests that the amount of speech production varies largely according to learner groups. Among four learner groups, Japanese learners produce the least amount of speeches, which is interestingly in accordance with the result of analysis of the LINDSEI (Guilquin *et al.* 2010, p.23).

3.4 Data Distribution

After checking the appropriateness of the files in terms of length, amount of utterance, and content, all the sound files are manually transcribed. The use of special transcription codes is limited to basic ones such as [unclear] and [ph]. The former is used when the

speech is completely incomprehensible to the transcriber; and the latter is used when the transcriber is uncertain of the correctness of his or her phonetic transcriptions.

As mentioned above, many speech corpora, including the LINDSEI, have distributed only the transcriptions. The ICNALE-Spoken plans to distribute the sound files together with the transcripts. Needless to say, the distribution of sound data is highly beneficial to learner corpus studies, but corpus builders need to be cautious about protecting the speakers' privacy (McEnery & Hardie, 2012, pp. 60-66).

Therefore, with the purpose of distributing the sound data without the risk of identification of speakers, we have developed another original system, the ICNALE-Spoken Automatic Speech Morphing System (ASMS).

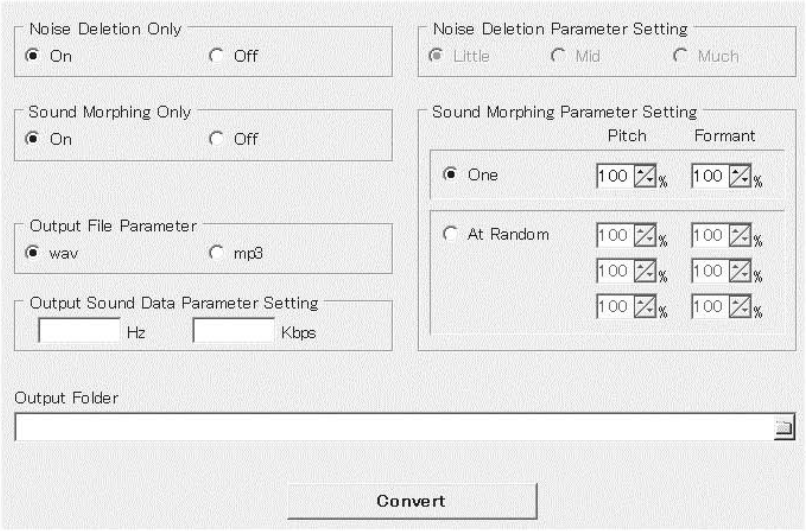

Fig. 6 Interface of the Beta Version of the ICNALE-Spoken Automatic Speech Morphing System (ASMS)

ASMS drastically changes the acoustic image of the original speaker's voice to the degree that hearers can never identify who the speaker is, even if they know him or her very well, by adjusting the pitch and formant of the original sound files.

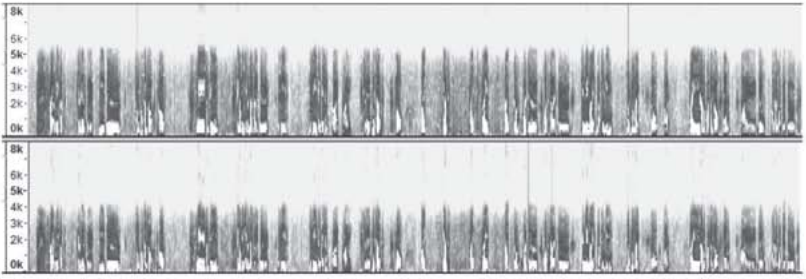

Fig. 7 The spectrograms before acoustic morphing with ASMS (upper) and after morphing (lower). Reproduced from Ishikawa (2014).

Fig. 8 The sound waves before acoustic morphing with ASMS (upper) and after morphing (lower). Reproduced from Ishikawa (2014).

Sound morphing changes the pitch and the formant, but not the sound wave itself, meaning that corpus users can conduct a basic acoustic analysis even with the morphed sound data. This innovation clearly widens the possibilities for analysing learners' spoken interlanguage productions immensely.

IV Conclusion

Although comparative analysis of spoken and written production is a promising field of research, previous corpus studies, of both learners and general L1 speakers, have been attempted only in a rather limited way, mainly due to the lack of appropriate speech databases. This study takes into serious consideration the view that "the medium of communication itself does produce a distinction which is linguistically meaningful… thinking about corpora in terms of mode of production is not just a matter of different data collection and technical issues… it is, rather, linguistically a very real distinction." (McEnery & Hardie, 2012, p.5)

In the current paper, we have introduced the outline of the ICNALE-Spoken with

comparative reference to the LINDSEI. As a new database enabling a reliable spoken-written, multi-modal, internationally contrastive interlanguage analysis, it is expected to contribute to the advancement of learner corpus studies.

References

Edwards, J. A. (1992). Design principles in the transcription of spoken discourse. In J. Svartvik (Ed.) Directions in corpus linguistics (pp. 129-147). Berlin & New York: Mouton de Gruyter.

Gilquin, G., De Cock, S., & Granger, S. (2010). *Louvain international database of spoken English interlanguage (LINDSEI).* Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37-51). Lund, Sweden: Lund University Press.

Granger, S., Dagneaux, E., & Meunier, F. (Eds.) (2002). *The international corpus of learner English: Handbook and CD-ROM.* Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.) (2009). *The international corpus of learner English: Handbook and CD-ROM.* Version 2. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world Vol. 1* (pp. 91-118). Kobe, Japan: Kobe University.

Ishikawa, S. (2014). Maruchimodaru-gata chukan gengo taisho bunseki shiryo to shiteno eigo gakushusha hanashikotoba kopasu no kaihatsu: ICNALE Spoken purojiekuto no nerai to gaiyo. [Development of a new spoken learner corpus for multi-modal contrastive interlanguage analysis: The aim and outline of the ICNALE spoken project]. *LET Kyushu-Okinawa Bulletin*, 14.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice.* Cambridge, UK: Cambridge University Press.

Appendix

Instruction for the ICNALE Automatic Speech Collection System (Version 201404)

Inst: *Welcome to the ICNALE speech data collection system. Now, kindly respond to the following ten questions. This recording will last for approximately ten minutes. If you stop performing the tasks halfway, you will have to do them all over again from*

the beginning. You are advised to complete all the tasks within a single session.

Q1: *After the beep, please key in your student number provided by your college using the keys on your phone.*

Q2: *Please state your family name and first name after the beep.*

Q3: *Please state the name of your nation or country.*

Q4: *Please state the name of your college or university after the beep.*

Inst: *If this is the first time for you to take this telephone interview, press 0. If not, press 1.*

Q5: *Please record your self-introduction after the beep. You may talk about your hobbies, your academic major, and your dreams; you may choose any topic you like. You have to continue talking for 60 seconds.*

Q6: *You may begin with speech task 1. Now, listen to the topic carefully. These days, some people say that XXX (The topic will be given in the actual interview). Do you agree or disagree with this statement? Use reasons and specific details to support your claim. You will have 20 seconds to prepare your response. After the beep, start talking immediately and continue for 60 seconds. Do not stop talking before the time is up. Do you like to listen to the topic once more again? If yes, press 1. If no, press 0 and prepare for your speech.*

Q7: *This is your second trial for your speech on XXX. You will have 10 seconds to prepare. You can repeat the same points, but this time, try to speak more. After the beep, start talking immediately.*

Q8: *You may begin with speech task 2. Now, listen to the topic carefully. These days, some people say that YYY (The topic will be given in the actual interview). Do you agree or disagree with this statement? Use reasons and specific details to support your claim. You will have 20 seconds to prepare. After the beep, start talking immediately. Do not stop talking before the time is up. Do you like to listen to the topic once more again? If yes, press 1. If no, press 0 and prepare for your speech.*

Q9: *This is your second trial for your speech on YYY. You will have 10 seconds to prepare. You can repeat the same points, but this time, try to speak more. After the beep, start talking immediately.*

Q10: *This is the last question. How was your speech today? Press the number on your phone after the beep. Very well -- press 3 Well ---- press 2 So-so --- press 1 Bad --- press 0.*

Inst: *Now you have completed all the tasks. Your task completion code is X X X X. Don't forget to fill the blank on your Excel Sheet with this code. Thank you very much for your cooperation for the ICNALE project.*