The NICT JLE Corpus: the final report

Yukio Tono
Meikai University
Emi Izumi
National Institute of Information and Communications Technology
Emiko Kaneko
ALC Press

Reference Data:
Tono, Y., Izumi, E., & Kaneko, E. (2005). The NICT JLE Corpus: the final report. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2004 Conference Proceedings*. Tokyo: JALT

This is the final report of the NICT JLE Corpus project. It is a project of compiling a spoken learner corpus based on the Standard Speaking Test developed by ALC Press. More than 1200 Japanese subjects' 15-minute interview data was transcribed and included in the corpus with various learner profile information and some linguistic annotations such as error tagging, fillers, self-correction, among others. The size of the corpus is approximately two million words, which is the biggest spoken learner corpus ever made in the world. This paper will report on the basic design criteria of the corpus, the process of its development and show its potential by presenting some preliminary research results using this corpus data.

本稿は日本人英語学習者１２００人余りのスピーキングデータを電子化したNICT JLE Corpus の公開にあたっての最終報告である。会話の英語学習者コーパスとして世界最大であるばかりでなく，Standard Speaking Test という会話テストの成績によりレベル分けされたデータであるため英語力の伸びと英語使用の特徴を関連付けて分析できる。それだけでなくエラーや談話関連の言語注釈も施されているので，さまざまな活用の可能性がある。本稿では，その開発の意図，経緯，関連ツールの紹介，そして完成したコーパスを用いた予備的な調査結果をもとにそのコーパス構築の意義と活用について論じたい。

There is a growing awareness that language teaching and learning should be better informed by electronic resources such as language corpora, electronic dictionaries and materials on the Internet among others. Especially the potential of the use of corpora for language learning has been gaining much attention as more products become available using corpora. Major monolingual learner's dictionaries, for example, so-called 'Big 4' (*LDOCE*, *OALD*, *CALD*, and *COBUILD*) as well as more recent *MED* (*Macmillan's English Dictionary*), all claim that they are 'fully corpus-based'. The first corpus-based TV conversation program, *100-go de start eikaiwa* (*Let's start with the first 100 words in English*) has been welcomed as a new type of lexically-oriented language syllabus.

The NICT JLE (Japanese Learner English Corpus): Overview

In this section, we will give an overview of the NICT JLE Corpus, mainly by explaining the nature of the SST (Standard Speaking Test) which is the source of the corpus data and the method by which learner data has been collected, transcribed, and annotated including error tagging. We will also mention two subcorpora which have been compiled in order to observe learners' language from a broad perspective.

Recording & Transcribing the Speech Data

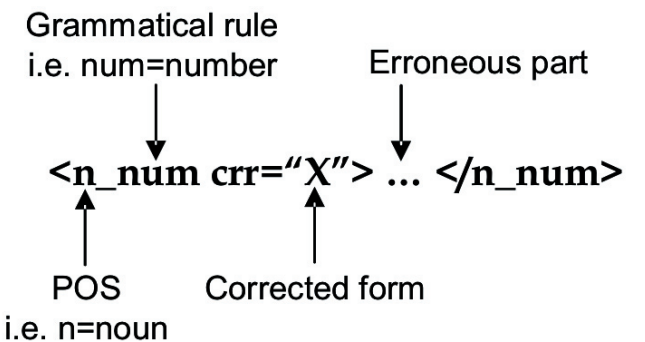
Each interview was recorded in a quiet room by means of DAT (Digital-Audio Tape) as the medium. There are some general rules for transcribing. For instance, even though a word may be mispronounced, it is transcribed with the correct spelling, provide that the transcribers are able to understand the word that was produced. If acronyms are pronounced as sequences of letters, they must be transcribed as a series of upper case letters, which are separated by spaces. Roman or Arabic numerals must not be used. All numbers must be transliterated as words. The transcribers are allowed to insert phrase and sentence boundaries with commas and periods, based on their own discretion. Some information on non-verbal behaviours or concurrent events such as relevant noises is also inserted.

Discourse Tagging

There are more than 30 basic tags for identifying discourse phenomena in the utterances. These are divided into four groups: tags for representing the structure of the interview, tags for the interviewee's profile, tags for speaker turns, and tags for representing utterance phenomena such as fillers, repetitions, self-corrections, overlapping, and so on.

Error Tagging

Analyzing errors produced by learners is an efficient way of finding out the learners' stages of development and for deciding the most appropriate teaching method for them. We are aware that it is quite difficult to design a consistent and generic error tagset as the learners' errors extend across various linguistic areas. We need to have a robust error typology to accomplish this. We designed the original error tagset only for learners' grammatical and lexical errors, which are relatively easy to categorize, compared with other error types such as discourse errors or errors related to more communicative aspects of learners' language. The error tagset consists of 47 tags. As shown in Figure 1, an error tag contains three pieces of information: part of speech, a grammatical and lexical rule, and a corrected form.



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

Figure1. Structure of an error tag and an example of an error-tagged sentence

By referring to the corrected form indicated in an error tag, it is possible to obtain a corrected sentence just by converting erroneous parts into corrected equivalents.

Subcorpora

We have also compiled two subcorpora for comparison. One is a native English speakers' corpus and the other is a back-translation corpus. The native English speakers' corpus is considered to be quite useful for comparing the utterances of native speakers and Japanese learners. We were able to make this comparison by collecting the speech data of native speakers', conducting a similar type of interview to that of the SST. The back-translation corpus was compiled mainly by guessing what the learners intended to say in the interview, and then translating this into correct Japanese. With the back-translation corpus, we were able to study how L1 (Japanese) transfer interferes with second language acquisition, or the kinds of things which are difficult for Japanese learners to express in English. As stated above, we performed error tagging only for grammatical and lexical errors. These subcorpora may cover what we are unable to examine solely by error tagging. The structure of the entire corpus and the size of each dataset are described in Figure 2.

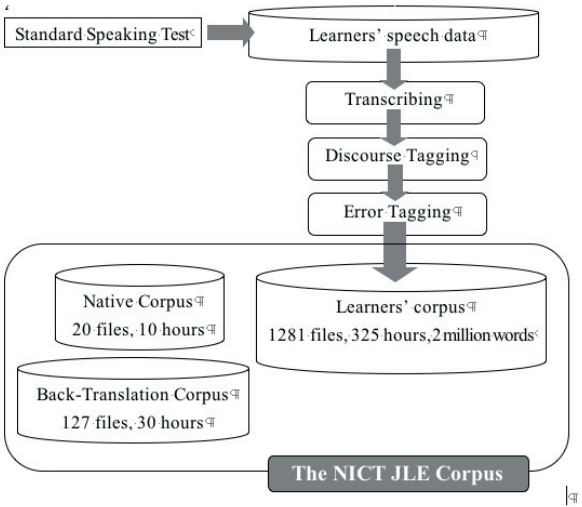


Figure 2. The Speech Corpus of Japanese Learners of English

What is the Standard Speaking Test?

The Standard Speaking Test (SST) was developed by ALC Press with ACTFL (American Council for the Teaching of Foreign Languages) in 1996. It was modelled after the ACTFL OPI, which, conducted in 37 languages, is also an interview test. Similarly, SST is a fifteen-minute, tape-recorded interview test, in which picture prompts and Role Play are always utilized in specific stages. Since its launch in 1997, more than 100 organizations have taken the SST.

Among many advantageous characteristics of the SST, the most striking is that it is an adaptive test. Specifically, SST is adaptive in terms of the levels and contents of the questions asked during the interview. Although the test is highly structured with the use of prompts, because of this characteristic no two interviews are the same, and it is impossible to prepare entirely scripted questions and answers.

There are nine levels in the SST, from Level One for Novice Low to Level Nine for Advanced, but five of these, Levels Four to Eight, cover the Intermediate Band. The test

is designed this way in order to meet the particular needs of Japanese ESL learners. The SST is holistically evaluated by at least two certified raters based on ACTFL Proficiency Guidelines. Raters evaluate interviews in terms of Global Tasks/Functions, Content/Context, Oral Text Type, and Accuracy. Inter-rater reliability as of July 2004 is 86.1%.

Study 1: Development of Noun Phrases in the Interlanguage of Japanese EFL Learners

Introduction

This research is a corpus-based study which seeks to analyze how Japanese EFL learners acquire noun phrases. The data for the research is extracted from the NICT JLE Corpus, a learners' spoken corpus, and its associated normative corpus. One of the distinguishing features of the NICT JLE Corpus is that the oral proficiency level of each examinee is included in the data, which makes it possible to observe, without conducting longitudinal research, how a certain grammatical feature develops as the oral proficiency level goes up. In the present study, this strength of the NICT JLE Corpus was fully utilized.

Hypotheses

Three hypotheses were made for this research.

- As the oral proficiency level goes up, the use of NPs with a postmodifier increases.
- The sequence of NP acquisition of Japanese EFL learners is as follows:
  - Adjective+Noun > Adverb+Adjective+Noun > Noun+Prepositional Phrase > Noun+Modifiying Clause

There is a correlation between oral proficiency levels and the frequency in which NPs with a postmodifier occur.

Method

Data

The data that were used for this research come from ten Intermediate High level speakers, ten Intermediate Mid level speakers, and ten Intermediate Low level speakers. For comparison, data from ten native speakers' were extracted from the normative corpus. The extracted data was POS tagged.

Table 1. Analyzed Data

SST Levels	Oral Proficiency Levels	Data Size
Native	Native Speakers (normative corpus)	4882 words
SST 8	Intermediate High (TOEIC avg. 874)	4857 words
SST 6, 7	Intermediate Mid (TOEIC avg. 812)	4020 words
SST 4, 5	Intermediate Low (TOEIC avg. 682)	3890 words

Instruments

Various types of NPs in the POS tagged data were searched with the use of monoconc. In order to verify the sequence of NP acquisition, SPSS was used to conduct a correspondence analysis. In doing so, NP types with a frequency less than 5 were eliminated. The cases in which an NP with a postmodifier occur were counted manually.

Results and Analysis

Hypothesis 1

Table 2 below shows that the use of postmodifiers increases and the use of premodifiers decreases in comparison with the expected frequency as the level of oral proficiency goes up. In fact, native speakers use more postmodifiers and fewer premodifiers than the expected frequency, which is a characteristic observed only in native speakers' data. This fact suggests that even Intermediate High level speakers haven't fully acquired using postmodifiers in the way native speakers do. Chi-square testing supports this result. The null hypothesis is false at 0.1% level of significance.

Hypothesis 2

Correspondence analysis was conducted to verify Hypothesis 2. In a correspondence map, the closer the distance between any two squares is, the more closely related they are. In this particular plotting, however, as Table 4 indicates, 91.4 % of the variance is explained by the first dimension, that is, in a horizontal distance.

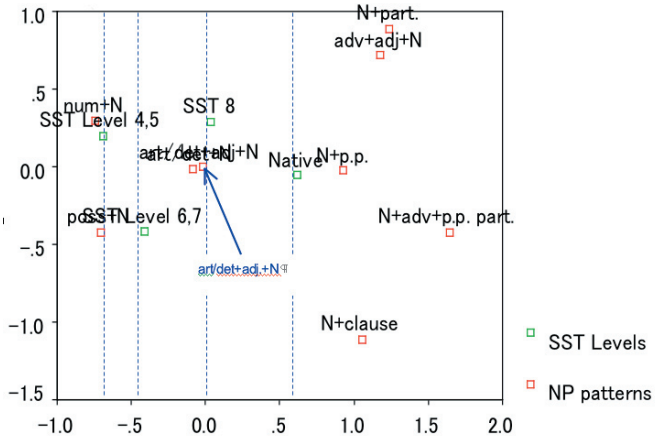


Figure 3. Correspondence Analysis Plots of Oral Proficiency Levels and NP Types

According to Figure 3, Intermediate Low is characterized by the simplest form of NP (numeral + noun, possessive pronoun + noun). Intermediate Mid, however, doesn't have any NPs which characterize this level. Presumably this is because there is a great variety in the profile of Intermediate Mid speakers, and no one NP type can characterize the level. Intermediate High is characterized by a relatively complex NP with a long premodifier (article/determiner +(adjective) + noun). Finally, native speakers are characterized by NPs with a long premodifier (adverb + adjective + noun) and prepositional phrases following the head NP.

The correspondence analysis suggests that the sequence of NP acquisition is from simple to more complex. However, it also indicates that Intermediate Mid performance cannot

Table 2: Modification Categories and the SST Levels

			SST Levels				TOTAL
			Level 4,5	Level 6,7	Level 8	Native	
Modification Category	PRE	Observed Frequency	297	308	322	419	1346
		Expected Frequency	264.8	287.8	322.8	470.6	1346.0
	POST	Observed Frequency	13	29	56	132	230
		Expected Frequency	45.2	49.2	55.2	80.4	230.0
TOTAL		Observed Frequency	310	337	378	551	176
		Expected Frequency	310.0	337.0	378.0	551.0	1576.0

Normalized with a corpus size of 5000 words

Table 3. Chi-square Test

	Value	DF	Asymp. Sig (2-sided)
Pearson's Chi-Square	75.365a	3	.000
Likelihood Ratio	80.505	3	.000
Linear-by-linear Association	73.563	1	.000
N of Valid cases	1576		

Table 4. Summary

Dimension	Singular Value	Inertia	Chi-square	Sig	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
								2
1	.263	.069			.914	.914	.022	-.151
2	.067	.004			.059	.973	.021	
3	.045	.002			.027	1.000		
TOTAL		.076	119.090	.000a	1.000	1.000		

a.33 Degree of Freedom

Table 5. Frequency of Postmodifiers in Each Oral Proficiency Level

	Native	Intermediate High	Intermediate Mid	Intermediate Low
Total # of NP	667	566	568	645
w/ post modifier	132	56	29	13

be characterized just by the use of particular NPs, which suggests that the sequence of acquisition may not be straightforward.

Hypothesis 3

The study about the cases in which long NPs occur has revealed that in native speakers' data, 67% of NPs with a postmodifier are used in the slot immediately following "There is/are" and as an object of a preposition, which is usually placed at the end of a sentence. They do not occur in the objective case as frequently as one would expect

Figure 4 shows that the distribution of NPs with a postmodifier is obviously skewed in the native speaker sample. As the level goes down, the distribution becomes more even, and the patterns found in native speaker data sre not observed in Intermediate Mid. Because the occurrence of postmodifiers itself is marginal in Intermediate Low as Table 5 shows, I disregard this level.

Summary and Implication

The three hypotheses, which no English teachers would feel an objection to, are confirmed to some extent. As the oral proficiency level goes up, the use of NPs with a postmodifier increases. Learners acquire a simple NP first and then

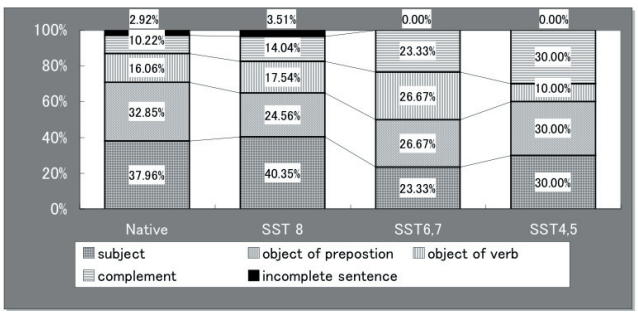


Figure 4. NP with a Postmodifier and their Cases

learn to use more complex ones. Interestingly, the study also indicates that the process of Japanese EFL learners' acquiring complex NPs is not straightforward. There is a preference in the cases where native speakers use NPs with a postmodifier. As the level of oral proficiency goes down, however, the preference disappears.

This is a preliminary study to show what can be done with the NICT JLE corpus. Depending on how researchers use it, the NICT JLE Corpus will have significant implications for English education and the field of SLA.

Study 2: Automatic Error Detection

In the support system for language learning, we have assumed that learners should be told what kind of errors they have made, and in which part of their utterances. To do this, we need to have a framework that will allow us to detect learners' errors automatically. In this section, we are going to demonstrate an experiment on automatic error detection in which we applied natural language processing (NLP) techniques by using error tag information in the NICT JLE Corpus. We will examine to what extent this could be accomplished using our learner corpus, by describing a method of detecting learners' grammatical and lexical errors and using other techniques that improve the accuracy of error detection with a limited amount of training data.

Method

Detection of Omission-type Errors

Omission-type errors are detected by determining whether or not a necessary word or expression is missing in front of each word, including delimiters (Figure 1, Method A). During this process, we also determined the category the error belonged to. The expression 'error categories' here means the 47 error categories that have been defined in our error tagset (e.g. article errors, tense errors, and so on). If more than one error category is given, we need to choose the most appropriate error category 'k' from among N+1 categories, which means we have added one more category (+1), namely 'There is no missing word.' (labelled with 'C') to the N error categories (Figure 5, Method B).

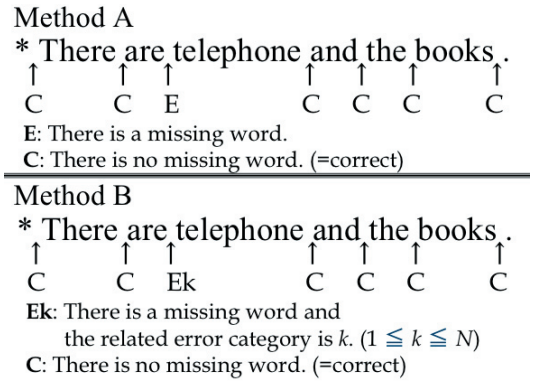


Figure 5. Detection of omission-type errors

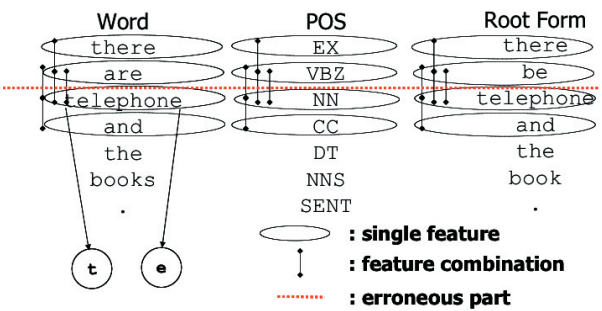


Figure 6. Features used for detecting omission-type errors

Method C

* I lived in the Japan in my childhood.

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

C C C Eb C C C C

Eb: The word at the beginning of the part which should be replaced.
Ee: The word in the middle or the end of the part which should be replaced.
C: No need to be replaced nor deleted. (=correct)

Method D

* I lived in the Japan in my childhood.

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

C C C Ebk C C C C

Ebk: The word at the beginning of the part which should be replaced and whose error category is *k*.
Ee: The word in the middle or the end of the part which should be replaced and whose error category is *k*. ($1 \leqq k \leqq N$)
C: No need to be replaced nor deleted. (=correct)

Figure 7. Detection of replacement/insertion-type errors

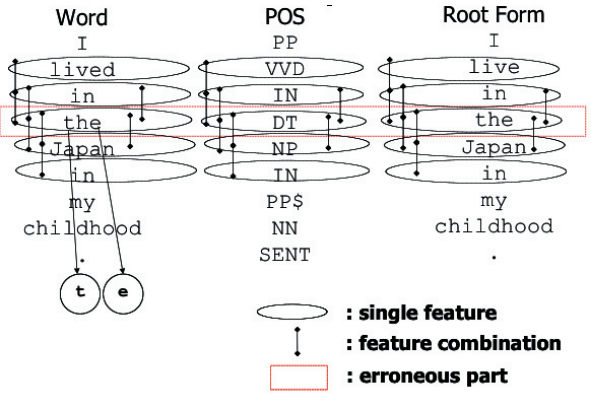


Figure 8. Features used for detecting replacement/ insertion type errors

To perform the estimation, we refer to 23 pieces of information as described in Figure 6. The word classes and root forms are obtained using 'TreeTagger' (Schmid 1994).

Detection of Replacement-type/Insertion-type Errors

Replacement-type and insertion-type errors are detected by estimating whether or not each word should be deleted or replaced with another word string. The error category is also determined during this process. If more than one error category is determined, we use two methods of detection, as shown in Figure 7. In Method C, if the word is to be replaced, the model estimates whether the word is located at the beginning, middle, or end of the erroneous part. Method D is used if N error categories arise. We choose an error category for the word from among 2N+1 categories. '2N+1 categories' means that we divide N categories into two groups, i.e., firstly when the word is at the beginning of the erroneous part and secondly when the word is not at the beginning. We add one more (+1) when the word neither needs to be deleted nor replaced. To do this, we applied Ramshaw's IOB scheme (Ramshaw and Mitchell, 1995).

To estimate an error category, we refer to 32 pieces of information, as shown in Figure 8.

Use of Machine Learning Model

We considered error detection as similar to class categorization, in which goal is, according to Manning and Schutze (1999), to classify the topic or theme of a document. Our first attempt is to apply machine learning model to our

framework. We chose the Maximum Entropy (ME) (Jaynes, 1957, 1979) model, which is used variously to solve class categorization problem and which is one of the general techniques for estimating probability distributions of data.

Experiment

Targeted Error Categories

As shown in Table 6, we selected 13 error categories for detection. We assume that these errors are more frequent than other errors, and can be identified relatively easily from the context.

Table 6. Error categories to be detected

Noun	Number error, Lexical error
Verb	Erroneous subject-verb agreement, Tense error, Compliment error, Lexical error
Adjective	Lexical error
Adverb	Lexical error
Preposition	Lexical error on normal and dependent preposition
Article	Lexical error
Pronoun	Lexical error
Others	Collocation error

Experiment Based on Tagged Data

We obtained 167 error-tagged transcripts from the NICT JLE Corpus. We used 151 files as training data, and 16 files as test data.

We tried to detect each error category using the method described above. Since there were some error categories that could not be detected due to the lack of training data, the overall rate was inadequate (Table 7). The best results were obtained for article errors, which were the most frequently occurring errors, as shown in Table 8.

Table 7. Recall/precision for the detection of all errors

All errors		
Omission-type	Recall	96/277 * 100= 34.66 %
	Precision	96/169 * 100= 56.88 %
Replacement-/Insertion-type	Recall	37/647 * 100= 5.72 %
	Precision	37/183 * 100= 20.22 %

Table 8. Recall/precision for the detection of article errors

Article errors		
Omission-type	Recall	86/172 * 100= 50.00 %
	Precision	86/143 * 100= 60.14 %
Replacement-/Insertion-type	Recall	13/88 * 100= 14.77 %
	Precision	13/44 * 100= 29.55 %

We assumed that the results were inadequate because we did not have sufficient training data. To compensate for the lack of training data, we added the correct sentences and the sentences with aritificially-made article errors to see how this would affect the results.

Addition of Correct Sentences and Sentences with Artificially-Made Errors

We added approximately 105,000 new correct sentences of the following three types: the native speakers' speech data subcorpus, the interviewers' utterances and the corrected sentences extracted from the error-tagged data. As for the sentences with the artificially-made article errors, article errors were automatically added by using simple manually-constructed rules. These rules were derived by investigating the characteristics of learners' errors found in our corpus. We first examined what kind of article errors had been made and found that there was often confusion between 'a', 'an', 'the' and the absence of an article. We made up pseudo-errors by replacing the correctly used articles with one of the alternatives. In total, approximately 7,600 new sentences that contained artificially made errors have been added to the training data. By doing this, the results have been improved as shown in Table 9 and 10. We obtained a better recall and precision rate for almost all types of errors.

Table 9: Recall/precision for the detection of all errors

All errors		
Omission-type	Recall	89/277 * 100= 32.13 %
	Precision	89/122 * 100= 72.95 %
Replacement-/Insertion-type	Recall	46/647 * 100= 7.11 %
	Precision	46/183 * 100= 25.14 %

Table 10: Recall/precision for the detection of article errors

Article errors		
Omission-type	Recall	89/172 * 100= 51.74 %
	Precision	89/116 * 100= 76.72 %
Replacement-/Insertion-type	Recall	19/88 * 100= 21.59 %
	Precision	19/25 * 100= 76.00 %

Summary

In this experiment, we have tried to detect learners' errors automatically based on the error-tagged information of the NICT JLE Corpus by using the machine learning technique. We found that adding the correct sentences or adding artificially-made errors, to the training data improves accuracy. However, to improve accuracy for the detection of replacement and insertion-type errors, we need to obtain more error-tagged sentences and examine global the context more thoroughly.

References

Aoi-Kimura, M. (2002). Describing Japanese EFL learners' process of noun phrase development: a learner corpus-based approach to interlanguage construction. *LEO*, 31: 11-34.

Jaynes, E. T. (1957) "Information theory and statistical mechanics", *Physical Review* 106: 620-630.

Jaynes, E. T. (1979) "Where do we stand on maximum entropy?" In R.D. Levine and M. Tribus (eds.), *The maximum entropy formalism*. MIT Press: 15.

Manning, C. D. and Schutze, H. (1999) *Foundation of Statistical Natural Language Processing*, MIT Press, pp. 575.

Ramshaw, L. A. and Mitchell P. M. (1995) "Text chunking using transformation-based learning", In *Proceedings of the Third ACL Workshop on Very Large Corpora*: 82-94.

Saito, T., Nakamura, J., Akano, I., (Ed.) . (1998). *Eigo corpus gengogaku kiso to jissen* [*English corpus linguistics: principles and practice*]. Tokyo: Kenkyusha.

Schmid, H. (1994) "Probabilistic part-of-speech tagging using decision trees", In *Proceedings of International Conference on New Methods in Language Processing*, pp. 45-49.

Tokyo Daigaku Kyoyogakubu Tokeikyousitu (1994). *Jinbunshakaikagakuno tokeigaku* [*Statistics.in humanities and social science*]. Tokyo: Tokyo University Press