English

Top

神戸大学　国際コミュニケーションセンター

石川慎一郎研究室

Dr. Shin Ishikawa, Kobe University

CEEAUS

科学研究費プロジェクト進捗報告ページ （2007～2009年度）
英語語彙力における意味・音韻知識の関係性解明と両者の統合を促す学習システムの開発（19720135）

Updated 2013/06/28
（本プロジェクトは2009年度で終了し，現在，このページの更新は終了しています。）

●はじめに

石川研究室では，2007～2009年度にかけて，日本人英語学習者の語彙使用を客観的に調査するための資料として，Corpus of English Essays Written by Asian University Students（CEEJUS）の構築を行ないました。このプロジェクトでは，日中韓３か国で約20万語の作文が収集されました。現在，CEEAUSのデータは，石川他（2010）『言語研究のための統計入門』（くろしお出版）に付属のCD-ROMでご覧いただくことができます。

CEEAUSの開発は2009年度で終了しましたが，2010年度からは，CEEAUSのコンセプトを継承するInternational Corpus Network of Asian Learners of English（ICNALE）という大規模国際コーパス開発プロジェクトが始まり，2012年度末に130万語のコーパスが一般公開されました。

●注意

下記は，CEEAUS公開時のリリースノート（2009年11月）です。なお，後継のICNALEではデータ収集基準が多くの点で変更されています。ICNALEの構築過程については下記をご参照ください。

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In Ishikawa, S. (Ed.), Learner Corpus Studies in Asia and the World Vol.1 (Kobe University), 91-118.

「アジア圏英語学習者コーパスCEEAUS」リリースノート

CEEAUS Release Note

石川　慎一郎（神戸大学）
Dr Shin Ishikawa (Kobe University)

リリースノート2版公開　2009/11/20
リリースノート初版公開　2008/12/1

1．学習者コーパスとは何か　What is a learner corpus?

一般に，コーパスと言えば，母語話者による母語産出を収集した言語データベースのことを指します。たとえば，イギリス英語1億語を収集したBritish National Corpusは，「英語」の信頼できる標本データとして，これまで，さまざまな辞書や教材の開発に使われてきました。しかし，最近では，第2言語（L2）学習者の言語産出を収集した学習者コーパス（learner corpus）にも関心が高まっています。時には間違いや不自然な言葉づかいを含む学習者のL2産出を見ることにどのような意味があるのでしょうか？

第2言語習得研究では，母語と目標言語の中間に存在する言語を「中間言語」（interlanguage）と呼びます。たとえば英語を学習中の日本人が書いたり話したりする英語は，日本語でもなく，かといって正しい英語でもない，つまりは，日本語と英語の中間に存在する言語というわけです。学習者コーパスは，こうした中間言語の研究，ひいては第2言語の習得メカニズムの研究に大きく役に立ちます。

また，母語話者コーパスに加えて学習者コーパスがあれば，母語話者の産出する言語と学習者の産出する言語（「中間言語」）を比較することができ，これにより，学習者ならではの間違いのパターンや，不自然な過剰・過少使用のパターンを特定することが可能になります。こうした比較研究の成果を教材開発や言語教育に導入していくことも重要です。

When we use the term 'corpus,' we usually refer to a language database produced by native speakers. For example, the British National Corpus, which includes 100 million words of written and spoken English produced by English (or rather "British") native speakers, has been widely used as a reliable sample of real English. Currently many dictionaries and TESOL materials are based on the BNC or similar native speaker corpora. However, increasing numbers of researchers have recently become interested in the "learner corpus" as well as the conventional native speaker corpus. Why do we need to examine learner English, which may include errors and awkward expressions?

In the study of SLA or second language acquisition, the special type of language existing between one's first language (L1) and second language (L2) is called interlanguage (IL). For

example, when Japanese learners of English speak or write in English, their oral or written utterance is defined as an IL. It is neither Japanese as their L1 nor native-like English as their L2, but a very unique type of language seen between them. Analysis of the learner corpus helps elucidate the features of the interlanguage and the psychological procedure involved in acquiring the second language.

Learner corpus studies are also beneficial for foreign language teaching. By comparing learner corpus with native speaker corpus, we can identify learners' typical errors and overused or underused patterns. Most of the recent ESL dictionaries compiled in UK utilize various findings obtained from learner corpus analysis.

---

2. 日本人英語学習者コーパスの開発　Various learner corpora

学習者コーパスの研究は主としてヨーロッパで広まりましたが，日本人英語学習者コーパスの構築や研究も進んでいます。Sylviane Granger氏が提唱する国際英語学習者コーパス（International Corpus of Learner English：ICLE）の第2版（2009）には，日本人モジュールが含まれています。また，東京外国語大学の投野由紀夫先生が開発されたJEFLL（Japanese EFL Learner）コーパス，名古屋大学の杉浦正利先生が開発されたNICE（Nagoya Interlanguage Corpus of English）も一般公開されています。前者は中高生レベルの英作文を，後者は大学生レベルの英作文を集めたもので，ともに高い学術的・教育的価値を持っています。

Various learner corpora have been compiled to date. The largest and most influential is the International Corpus of Learner English (ICLE) (Granger, et al. 2002; Granger, et al., 2009).

Also, several corpora focus exclusively on Japanese learners of English (JLE). The Japanese EFL Learner Corpus (JEFLL) (Tono, 2007) includes essays written by Japanese junior high and high school students, the Nagoya Interlanguage Corpus of English (NICE) (Sugiura, 2007) includes those by Japanese college students, and the NICT JLE Corpus (Izumi, Uchimoto, & Isahara, 2004) collects transcribed data of speech utterances in the English oral proficiency interview (OPI) test.

Each of the existing learner corpora has its own merits, but when using them for what Granger (2002) calls contrastive interlanguage analysis (CIA), namely, a comparison between native speakers (NS) and non-native speakers (NNS) or that between NNSs with different L1s, we must be careful about how we interpret the obtained findings, since differences in the writing condition might influence the language used in the essays (Sugiura, 2007). It goes without saying that it is quite difficult to compare an academic essay concerning the global economic depression written by an NS and a short casual essay about a summer vacation by an NNS and then discussing aspects of the NS-NNS gap, which might actually be just the topic gap.

---

3. 学習者コーパスの研究利用　Using learner corpora for research purposes

Granger（1998）は，対照中間言語分析（contrastive interlanguage analysis：CIA）の重要性を指摘しています。CIAには2種類あり，1つは母語話者（NS）と非母語話者（NNS）の比較，もう1つは異なる母語を持つNNS同士の比較です。Grangerを中心とするICLEプロジェクトメンバーは，こうした観点から多くの研究成果を上げてきました。Granger（1998）の邦訳である船城道雄・望月通子監訳『英語学習者コーパス入門：SLAとコーパス言語学の出会い』（研究社出版，2008）には，ICLEを用いたCIA研究の成果がわかりやすく紹介されています。
今後の学習者コーパス研究の方向を考える場合，CIAの多層的展開が1つの視点となるでしょう。石川（2010）では，多層的対照中間言語分析（multi-layered CIA：MCIA）を提唱しました。これは，上記の2種類の比較に加え，異なる習熟度のNNS間比較，異なる態度要因を持つNNS間比較，さらには，学習者による第2言語産出と第1言語産出の比較が含まれます。こうした研究目的にあわせて開発されたのが後述するCEEAUSです。

CEEAUS is intended to be used as a database for a multi-layered contrastive interlanguage analysis (MCIA) (Ishikawa, 2010). With CEEAUS, you can compare (i) Japanese learners of English and English native speakers, (ii) Chinese learners of English and English native speakers, (iii) Japanese learners of English and Chinese counterparts, and (iv) English essays and Japanese essays written by Japanese native speakers respectively.

---

4. CEEAUSの概要 Outline of the CEEAUS

CEEAUS（Corpus of English Essays Written by Asian University Students）とは，神戸大学石川研究室で開発した日中韓の英語学習者コーパスです。日本人英語学習者が書いた英語作文を集めたCEEJUS，中国人英語学習者が書いた英語作文を集めたCEECUS，英語母語話者が書いた英語作文を集めたCEENAS，それに，日本語母語話者が書いた日本語作文を集めたCJEJUSの4つのモジュールが公開されています。

The Corpus of English Essays Written by Asian University Students (CEEAUS) (Ishikawa, 2008; Ishikawa, forthcoming) is a new learner corpus compiled and released by the author. CEEAUS currently consists of five modules, the CEEJUS module collecting essays written by Japanese university students, the CEECUS module collecting those by Chinese counterparts, the CEENAS module collecting those by English NS, and CJEJUS module collecting Japanese essays written by Japanese university students.

モジュール構成 Composition of Modules

●日本人英語学習者コーパス CEEJUS（Corpus of English Essays Written by Japanese University Students）
●中国人英語学習者コーパス CEECUS（Corpus of English Essays Written by Chinese University Students）
●英語母語話者コーパス CEENAS（Corpus of English Essays Written by Native Speakers）
●日本語母語話者コーパス CJEJUS（Corpus of Japanese Essays Written by Japanese University Students）

コーパスのサイズは下記の通りです。The sizes of the subcorpora are shown in the table below:

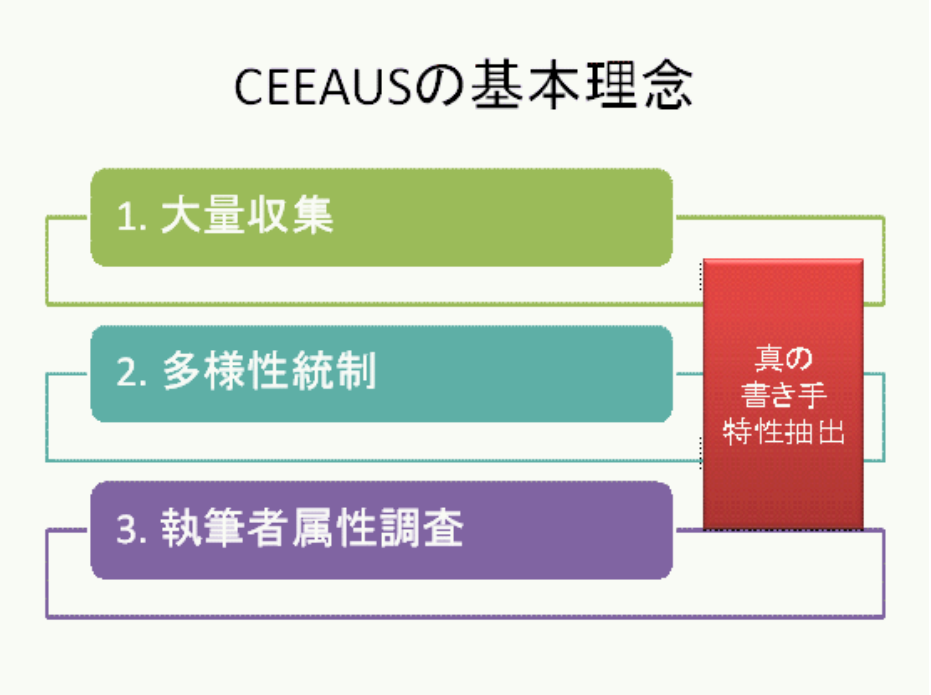
#	CEEJUS	CEECUS	CEENAS	CJEJUS
essays	770	92	146	50
token	169654	20367	37173	15344
type	4800	1818	3797	1653
lemma	3602	1472	2884	-----

---

5. CEEAUS開発の基本理念 Principles in the compilation of the CEEAUS

CEEAUS開発にあたっては，「大量収集」「多様性統制」「執筆者属性」の3つの点を基本理念としました。

There were three principles in compilation of the CEEAUS: large-scaled data collection, strict control of writing conditions, and demographic data collection.



まず，1については，前述のように全体で20万語を超えるデータを集めています。日本人以外の書き手については今後さらにデータ収集を行う必要がありますが，CEEJUSについては延べ770人の書き手による作文を集めており，既存のコーパスに比べても，かなり多くのデータを集めることができたと考えています。

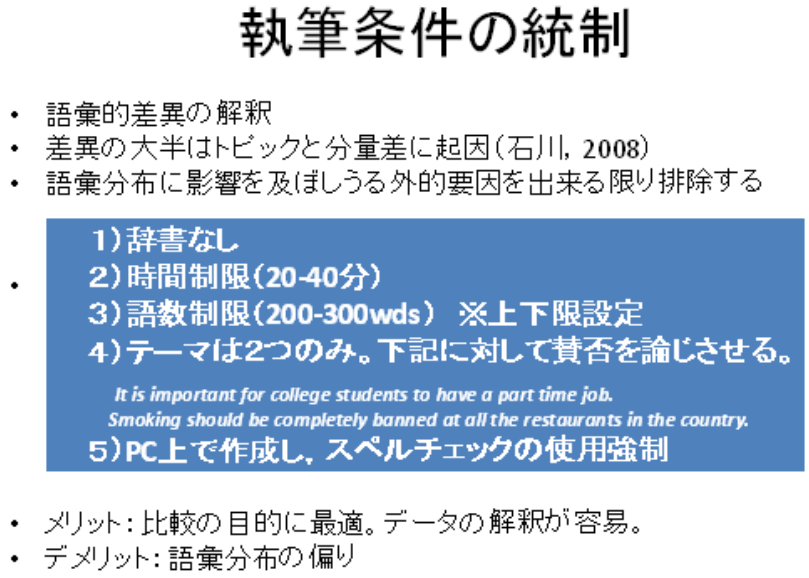
[1] Large-scaled data collection
CEEAUS has collected more than 200 thousand words in total. Though the volume of data of Chinese learners and English native speakers is relatively small. CEEJUS is one of the largest corpora ever compiled of Japanese learners of English at intermediate and advanced levels.

2については，従来の学習者コーパス研究では，CIAで得られた差分が書き手属性に起因するのか，それ以外のノイズ要因に起因するのかの切り分けが難しいという問題がありました。たとえば，ある単語が日本人英語学習者に典型的に多用されていることが分かったとしても，もしかしたらそれは，たまたま日本人英語学習者が多く選んだ作文テーマによるものであったかもしれません。同様に母語話者の特徴語を見つけたとしても，もしかしたら，それは母語話者のほうが長い作文を書いていたことによるものかもしれません。そこで，こうした解釈の曖昧さをできる限り解消するために，CEEAUSでは執筆条件の統制に意を砕きました。とくに，トピックを2種類（アルバイト：It is important for college students to have a part time job／禁煙：Smoking should be completely banned at all the restaurants in the country）に限定したことは，CEEAUSのユニークな特徴の一つです。

[2] Strict control of writing conditions
As mentioned before, control of writing conditions is vital when compiling a learner corpus. In CEEAUS, the number of topics is restricted to two.

(A) "It is important for college students to have a part time job."
(B) "Smoking should be completely banned at all the restaurants in the country."

Writers must clearly show whether they agree or disagree with the theses and support their own discussion with appropriate reasons and illustrative examples. Half the essays included in CEEAUS are written about topic A, and the remainder about topic B. In addition, the length of the essays (200 to 400 tokens) and the time allowed for writing (20 to 40 mins) are also controlled (Ishikawa, 2008). Dictionary use was not permitted.

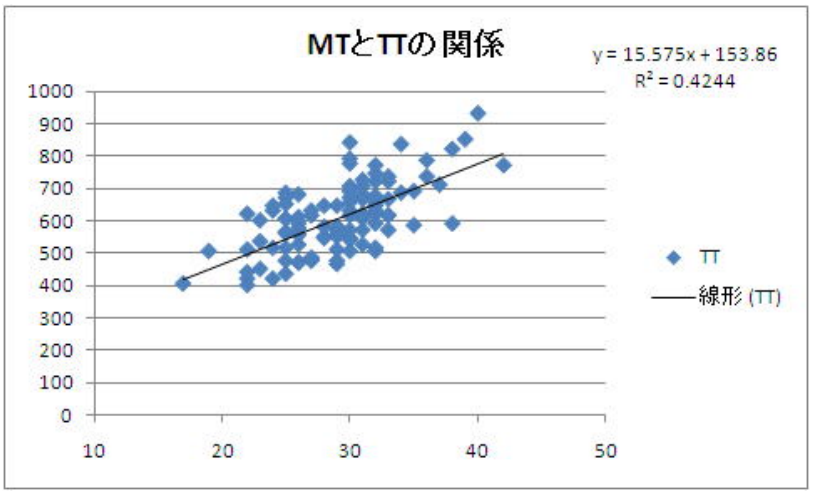


3については，書き手個々について，習熟度や多様な心理テストの結果などを収集しました。心理テストの結果は公開対象外にしていますが，習熟度については4段階に区分したデータを公開しています。具体的には，770人の全員に同じTOEIC(R)テスト型の模擬試験（MT）を受験させました。そして，770人のうち，実際にTOEIC(R)テスト（TT）を受験した学生150人のスコアとの対照を行い，回帰分析によって回帰式を導出し，模擬試験スコア（Mock Test）をTOEIC(R)推定スコアに変換しました。その上で，TOEIC(R)推定スコア700～をUpper，600～をSemi-upper，500～をMidlle，～495をLowerと決定しました。

[3] Demographic data collection
The data balance is carefully considered in CEEJUS as the main module of the same. All the writers took the TOEIC® test or mock TOEIC® test, and their essays are stratified into four L2 proficiency groups according to the actual or estimated test scores.

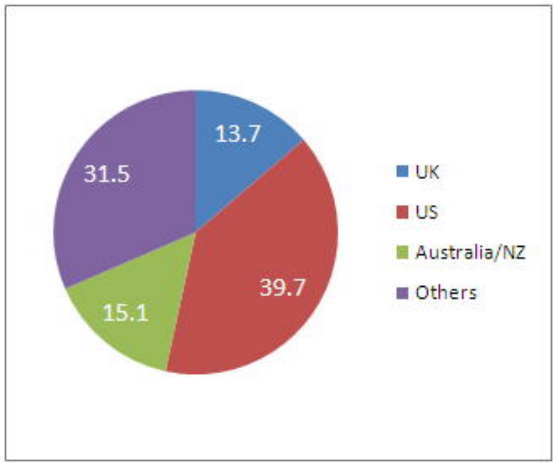
Upper (700+; 9,008 tokens in total)
Semi-upper (600+; 57,452 tokens)
Middle (500+; 85,614 tokens)
Lower (500–; 17,580 tokens)

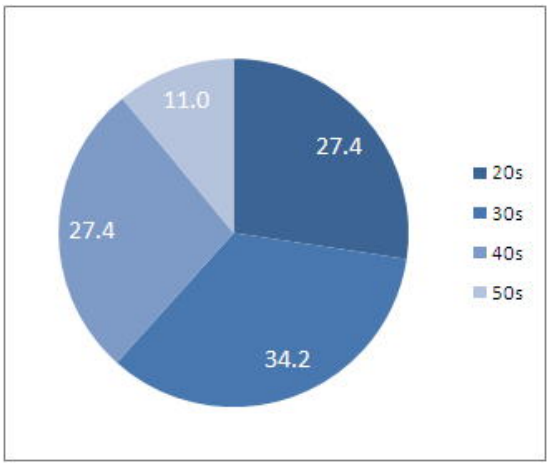
Thus, CEEJUS can also be used to compare non-native speakers at varying levels of English proficiency.



母語話者は，全員が成人で，78%が教師，残り21%がその他です。国籍・年齢の内訳は下記の通りです。

The demographic balance is also considered in CEENAS. Approx. 40% of the writers are American, and 15% each British and Australian. Age is also controlled: Approx. 35% of the writers are in their 30s and 28% in their 20s.





---

6 関連文献　References

Ishikawa, S. (2008). Eigo koupasu to eigo kyouiku: deeta toshiteno tekusuto. [English Corpus and Language Education: Text as a Data] Tokyo: Taishukanshoten.
Ishikawa, S. (2009). Vocabulary in interlanguage: A study on Corpus of English Essays Written by Asian University Students (CEEAUS). In K. Yagi & T. Kanzaki (Eds.) Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan (pp. 87-100). Nishinomiya, Japan: Kwansei Gakuin University Press.

If you have any questions, please ask the author.

Back to top