# 文法誤り情報および品詞／句情報付き英語学習者コーパスの構築

永田　亮†　Edward Whittaker††　Vera Sheinman††
† 甲南大学知能情報学部　†† 株式会社教育測定研究所
E-mail: †rnagata@ konan-u.ac.jp., ††{whittaker,sheinman}@jiem.co.jp

## 1. はじめに

関連分野における重要性にも関わらず，一般公開された学習者コーパスの数は限られている．特に，文法誤り情報が付与された学習者コーパスの公開数は非常に少ない．表1(注1)に示すように，文法誤り情報が付与されていたとしても，一般公開されていない，データへのアクセスに制限があるなど，その利用条件には制限が多い．例えば，文法誤りが付与された学習コーパスとして最大規模である Cambridge Learner Corpus は，Cambridge University Press の著者か Cambridge ESOL のスタッフに利用が限られている．

文法誤り情報が付与された学習者コーパスは，文法誤りの検出／訂正手法 [2]〜[4], [7], [8], [11]〜[14], [16], [17] の研究に必要不可欠である．しかしながら，上述の通り，学習者コーパスの公開は限られているため，研究者は独自のコーパスを構築し，手法の考案と評価を行っている．言い換えると，研究者間で共通の評価用データが存在しない．そのため，各手法の性能比較を行うことが困難であるという問題が生じている．したがって，文法誤り情報付きの学習者コーパスの公開は，同分野の更なる発展に大きく寄与する．

更に，形態素情報や句情報などの文法情報が人手で付与された学習者コーパスは，我々が知る限り公開されていない．このことは，次のような疑問に繋がる：“既存の解析技術は学習者の英文に対して上手く動作するのであろうか？” 現状では，この疑問に関する調査を行った研究は数少ない．Tetreault ら [17] は，学習者の英文に対する解析性能を一部調査しているが，対象としているのは前置詞のみである．学習者の英文は，綴り誤り，文法誤り，不自然な表現など，既存の解析技術が想定していない言語現象を多く含む．そのため，解析性能が低下すると予想される．それにもかかわらず，多くの研究者は既存技術を学習者の英文の解析に用いている．例えば，誤り検出／訂正手法では，処理過程で，品詞解析や句解析を行うのが一般的である．そのため，解析ミスが，誤検出や誤訂正の大きな要因となる可能性もある．また，コーパス言語学では解析技術を利用して，学習者コーパスから特徴的なパターンや表現を抽出する試みが盛んに行われている（例えば，文献 [1], [5], [18]）．解析ミスにより，誤った結論を導き出す可能性がある．

このような背景を受けて，我々は，研究教育活動に利用可能な学習者コーパス “Konan-JIEM Learner Corpus （KJ コーパス）” を構築した．同コーパスでは，文法誤り情報と文法情報（品詞情報と句情報）が人手で付与されている（図1に英文例を示す）．本稿では，この学習者コーパスの詳細を報告する．また，誤り情報と文法情報の付与に必要となるガイドラインについても述べる．なお，本学習者コーパスを言語資源協会（http://www.gsk.or.jp/）より公開している．

(注1)：Availability 欄において Yes はテキストデータへ直接アクセス可能なことを示す．Partially は，テキストデータへのアクセスが制限されていることを示す (例えば，専用のインタフェイスを通してのアクセスのみなど)．

文法誤り情報付き英文

<at crr="The"></at> Seasun <v_tns crr="was">is</v_tns> winter.
It <v_tns crr="was">is</v_tns> very cold.
I <v_tns crr="played">play</v_tns> skky and
<v_tnscrr="played">play</v_tns> <prp crr="with"></prp> friends.
It <v_tns crr="was">is</v_tns> very interesting.
I <v_tns crr="have">had</v_tns> a good memory.

文法誤り情報付き英文

[NP Seasun/NN ] [VP is/VBZ ] [NP winter/NN ] ./.
[NP It/PRP ] [VP is/VBZ ] [ADJP very/RB cold/JJ ] ./.
[NP I/PRP ] [VP play/VBP ] [NP skky/NN ] and/CC [VP play/VBP ] [NP friends/NNS ] ./.
[NP It/PRP ] [VP is/VBZ ] [ADJP very/RB interesting/JJ ] ./.
[NP I/PRP ] [VP had/VBD ] [NP a/DT good/JJ memory/NN ] ./.

図1：　KJ コーパス中の英文例

## 2. Konan-JIEM Learner Corpus

### 2.1 コーパスの概要

KJ コーパスに収録されている英文は，文献 [15] の研究を通じて収集したものである．英文の内容は，大学生がブログシステム上で書いた英文エッセイである．ブログシステムを用いたのは，英文の収集と管理を容易にするためである．ブログシステム上で，通常のエッセイライティングを行ってもらい，ブログ特有の機能などは使用していない．また，ライティングの際には，他人が書いたエッセイや自分自身が以前に書いたエッセイにはアクセスできないようにした．平均して週1〜2回のエッセイライティングを行い，最大で1人あたり10エッセイを書いてもらった．トピックは表2に示す通りである．

表 1： 学習者コーパスリスト

| Name | Error-tagged | Parsed | Size (words) | Availability |
|---|---|---|---|---|
| Cambridge Learner Corpus | Yes | No | 30 million | No |
| CLEC Corpus | Yes | No | 1 million | Partially |
| ETLC Corpus | Partially | No | 2 million | ? |
| HKUST Corpus | Yes | No | 30 million | No |
| ICLE Corpus [6] | No | No | 3.7 million+ | Yes |
| JEFLL Corpus [18] | No | No | 1 million | Partially |
| Longman Learners' Corpus | No | No | 10 million | ? |
| NICT JLE Corpus [9] | Partially | No | 2 million | Yes |
| Polish Learner English Corpus | No | No | 0.5 million | No |
| Janus Pannoius University Learner Corpus | No | No | 0.4 million | ? |

表 2： エッセイトピック

| Topic ID | Topic |
|---|---|
| 1 | University life |
| 2 | Summer vacation |
| 3 | Gardening |
| 4 | My hobby |
| 5 | My frightening experience |
| 6 | Reading |
| 7 | My home town |
| 8 | Traveling |
| 9 | My favorite thing |
| 10 | Cooking |

表 3： KJ コーパスの概要

| | |
|---|---|
| Number of essays | 233 |
| Number of writers | 25 |
| Number of words | 25,537 |

表 3 に，KJ コーパスの概要を示す．現状では，233 エッセイのうち 170 エッセイを一般に公開している．将来的には，全てのエッセイを公開する予定である．

### 2.2 文法誤り情報の付与

KJ コーパスでは，文法誤り情報付与のためのガイドラインの基礎として，NICT JLE コーパス [9] のガイドラインを使用することとした（NICT JLE コーパスのガイドラインの詳細は文献 [10] などを参照のこと）．同ガイドラインでは，XML 形式で，誤り箇所，誤りの種類，訂正情報のタグ付けを行う．例えば，図 1 中の “It <v_tns crr=”was”>is</v_tns> very cold.” は，タグ付けされた “is” に動詞の時制誤りがあり，“was” が正しいことを意味する．ここで，v_tns と crr=”was” は，それぞれ，動詞の時制誤りと訂正情報を表す．

変更点のひとつとして，フラグメントの扱いがある．フラグメントとは文が途中で終了している誤りのことである（例：“I have many books. Because I like reading.”）．KJ コーパスでは，フラグメントをタグ付けするため，新たなタグ <f> を定義した．例えば，“I have many books. <f>Because I like reading.</f>” のように，フラグメントにあたる部分に同タグを付与する．

また，NICT JLE コーパスのガイドラインで定義されている 46 種類の誤りタグを 22 種類まで減少させた（付録 1. に，タグの一覧を記す．詳細については，コーパスと共に公開しているガイドラインを参照のこと）．タグの種類が多いほど詳細に文法誤りの情報を付与できる．一方で，タグ付けに要する労力は増し，タグ付けの一貫性を保つことが難しくなる．すなわち，タグの種類数（情報の量）とタグ付けの難しさはトレードオフの関係にある．KJ コーパスでは，タグ付けの容易さと正確さを重視し，22 種類までタグ数を減らした．削除されたタグは，残されたタグに統合した．例えば，NICT JLE コーパスのガイドラインでは，名詞に関する誤りとして 6 種類のタグ（inflection, number, case, countability, complement, lexis）が用意されている．一方，KJ コーパスでは 3 種類である（number, lexis, other）．削除されたタグは，名詞その他の誤りを表す <n_o> に統合した．

### 2.3 文法情報の付与

KJ コーパスには，品詞情報と句情報も付与されている．ガイドラインの基礎として Penn Treebank のガイドラインを選択した．Penn Treebank のガイドラインは，学習者の書いた英文を想定していないため，次の 3 点について変更を行った（詳細については，コーパスと共に公開しているガイドラインを参照のこと）：

（1） 空白の抜け
（2） 文法誤り
（3） 綴誤り

（1）空白の抜けとは，トークン間の空白が抜けている誤りのことである．具体例として，“Tonight,we” や “beautifulhouse” などが挙げられる．抜けている空白を補い，通常どおりの手順でタグ付けを行う方法も考えられるが，空白の抜けに関する情報を保持するため KJ コーパスでは別の方法を選択した．補助的なタグとしてハイフン（-）を用い

る．空白の抜けがない場合に各トークンに付与されるタグをハイフンでつないで付与する．例えば，先ほどの例の場合，“Tonight,we/NN-,-PRP” と “beautifulhouse/JJ-NN” とタグ付けする．句情報の場合も同様に，句に対応するタグをハイフンでつないで付与する．例えば，“Tonight,we” は “[NP-PH-NP Tonight,we/NN-,-PRP ]” となる．ただし，タグ PH は，通常句としてタグ付けされないトークンを表す（cf., [NP Tonight/NN ] ,/, [NP we/PRP ]）．

（2）文法誤りに関しては，文法誤りの有無にかかわらず，対象トークンの表層情報に基づいて品詞／句情報をタグ付けすることを基本方針とした．例として，“There is apples.” を考える．この場合，訂正候補として，少くとも “There are apples.”，“There is an apple.”，“There is apple.” の三種類が考えられるが，本ガイドラインでは，各トークンの表層情報に基づき “[NP There/EX ] [VP is/VBZ ] [NP apples/NNS ] ./.” とタグ付けする．もし，文法誤りにより（もしくは学習者の英文に特有な他の言語現象により），付与すべきタグが不明な場合は，品詞に関しては UK を，句に関しては XP をそれぞれ使用する．

更に，追加仕様として CE タグを定義した．CE タグは，あるトークンが通常は使用されない品詞で使用されている時に使用する．例えば，“I don't success cooking.” では，通常動詞ととならない success が動詞の文脈で使用されている．この場合，CE タグを用いて，表層情報から決定される品詞（NN）と誤りがないときに与えられる品詞（言い換えると，文脈から決定される品詞 VB）の両方をコーディングする．例えば，上述の例では，I don't success/CE:NN:VB cooking. となる．すなわち，CE タグと 2 種類の品詞をコロンで連結してタグ付けする．これにより，利用者は目的に応じて必要な品詞を選択できる．ただし，CE タグは，“対象トークンの表層情報に基づいてタグ付けを行う” という基本方針に準拠していることに注意する必要がある．なぜなら，CE タグ，コロン，最後の品詞タグを削除することで基本方針のタグ付け結果が得られるからである（i.e., succes/NN）．

（3）綴誤りに関しては，正しい綴が推測できる場合には，正しい綴の語に付与される品詞タグと句タグを使用する（e.g., [NP sird/JJ year/NN ]）．推測できない場合は，UK と XP を使用する．

## 3. おわりに

本稿では，文法誤り情報と品詞／句情報を人手で付与した英語学習者コーパスである Konan-JIEM Learner Corpus（KJ コーパス）について報告した．KJ コーパスは，文法誤り検出／訂正手法の開発，学習者の英文を対象とした解析技術の開発，第二言語習得に関する研究などの分野において，重要なデータとなることが期待される．今後は，規模の拡大とともに，付与する情報を充実させることを予定している．

## 付　　録

### 1. 文法誤り情報付与のためのタグセット

文法誤りの情報を付与するためのタグセットの一覧である．このタグセットは，NICT JLE コーパスで使用されているタグセット [10] を基本に作成した．

- n: noun
  - num: number
  - lxc: lexis
  - o: other
- v: verb
  - agr: agreement
  - tns: tense
  - lxc: lexis
  - o: other
- mo: auxiliary verb
- aj: adjective
  - lxc: lexis
  - o: other
- av: adverb
- prp: preposition
  - lxc: lexis
  - o: other
- at: article
- pn: pronoun
- con: conjunction
- rel: relative clause
- itr: interrogative
- olxc: errors in lexis in more than two words
- ord: word order
- uk: unknown error
- f: fragment error

## 参 考 文 献


[1] J. Aarts and S. Granger, Tag sequences in learner corpora: a key to interlanguage grammar and discourse, Longman Pub Group, London, 1998.

[2] M. Chodorow and C. Leacock, “An unsupervised method for detecting grammatical errors,” Proc. of 1st Meeting of the North America Chapter of ACL, pp.140–147, 2000.

[3] M. Chodorow, J.R. Tetreault, and N.R. Han, “Detection of grammatical errors involving prepositions,” Proc. of 4th ACL-SIGSEM Workshop on Prepositions, pp.25–30, 2007.

[4] R.D. Felice and S.G. Pulman, “A classifier-based approach to preposition and determiner error correction in L2 English,” Proc. of 22nd International Conference on Computational Linguistics, pp.169–176, 2008.

[5] S. Granger, “Prefabricated patterns in advanced EFL writing: collocations and formulae,” in Phraseology: theory, analysis, and application, ed. A.P. Cowie, pp.145–160, Clarendon Press, 1998.

[6] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot, International Corpus of Learner English v2, Presses universitaires de Louvain, 2009.
[7] N.R. Han, M. Chodorow, and C. Leacock, "Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus," Proc. of 4th International Conference on Language Resources and Evaluation, pp.1625–1628, 2004.
[8] N.R. Han, M. Chodorow, and C. Leacock, "Detecting errors in English article usage by non-native speakers," Natural Language Engineering, vol.12, no.2, pp.115–129, 2006.
[9] E. Izumi, T. Saiga, T. Supnithi, K. Uchimoto, and H. Isahara, "The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques," Proc. of the Corpus Linguistics 2003 Coference, pp.359–366, 2003.
[10] E. Izumi, K. Uchimoto, and H. Isahara, "Error annotation for corpus of japanese learner english," Proc. of 6th International Workshop on Linguistically Annotated Corpora, pp.71–80, 2005.
[11] E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara, "Automatic error detection in the Japanese learners' English spoken data," Proc. of 41st Annual Meeting of ACL, pp.145–148, 2003.
[12] J. Lee and S. Seneff, "Correcting misuse of verb forms," Proc. of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technology Conference, pp.174–182, 2008.
[13] R. Nagata, A. Kawai, K. Morihiro, and N. Isu, "A feedback-augmented method for detecting errors in the writing of learners of English," Proc. of 44th Annual Meeting of ACL, pp.241–248, 2006.
[14] R. Nagata, F. Masui, A. Kawai, and N. Isu, "Recognizing article errors based on the three head words," Proc. of Cognition and Exploratory Learning in Digital Age, pp.184–191, 2004.
[15] R. Nagata and K. Nakatani, "Evaluating performance of grammatical error detection to maximize learning effect," Proc. of 23rd International Conference on Computational Linguistics, poster volume, pp.894–900, 2010.
[16] R. Nagata, T. Wakana, F. Masui, A. Kawai, and N. Isu, "Detecting article errors based on the mass count distinction," Proc. of 2nd International Joint Conference on Natural Language Processing, pp.815–826, 2005.
[17] J. Tetreault, J. Foster, and M. Chodorow, "Using parse features for preposition selection and error detection," Proc. of 48nd Annual Meeting of the Association for Computational Linguistics Short Papers, pp.353–358, 2010.
[18] Y. Tono, "A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora," Practical Applications in Language Corpora, pp.123–132, 2000.