# 解説文生成研究のためのライティング技術解説付き学習者コーパス

永田　亮†,††,†††　石川慎一郎††††　乾　健太郎†††††,†††
† 甲南大学　†† 国立研究開発法人科学技術振興機構，さきがけ
††† 理化学研究所 AIP センター　†††† 神戸大学　††††† 東北大学
E-mail: †nagata-nlp2019@ hyogo-u.ac.jp., ††iskwshin@gmail.com, †††inui@ecei.tohoku.ac.jp

## 1. はじめに

本稿では，解説文生成技術の実現を目指して，我々が構築した解説文データについて報告する．解説文生成とは，与えられた文書に対してライティング技術に関する解説を生成するタスクのことである．また，解説文とは，書き手のライティング能力の向上につながるヒントや説明のことである．図 1 に示すように，典型的には，文法誤りや不自然な表現に対する解説であることが多い．このような解説文により，学習者は，どこが悪いのか，どのように改善したらよいのかを，背後にある規則と共に理解することができ，自身のライティングのヒントを得ることができる．

本データ構築の目的は，解説文生成に関する研究を促進させることにある．解説文生成技術が実現すれば，英語ライティング学習支援などに大きく資すると期待できる．しかしながら，2. で述べるように，その研究例は非常に少ない．主な要因として，研究利用可能な一般公開されたデータが存在しないことを挙げることができる．解説文が付与されたコーパスは，解説文生成の研究において必要不可欠である．

このような背景を受けて，我々は，既存の二種類の英語学習者コーパス（ICNALE [5] の written essays と KJ [12]）に，人手で解説文を付与した．作成に先立ち，解説文としてどのような事項を付与すべきかを検討し，アノテーションガイドラインを策定した．その結果，解説文の内容として，ライティング技術一般に関するものと前置詞誤りに特化したものの二種類の解説文を付与することとした．一般的な解説に加えて，前置詞誤りを含めたのは，誤り数が特に多く，また，誤用の箇所と修正内容がはっきりしており解説が比較的容易であるためである．現状で，それぞれ，延べ約 2,000 文書と約 1,300 文書に対して付与作業が完了している（将来的には，それぞれ 6,000 文書と 3,000 文書に対して付与を行う予定である）．この結果を Web で公開している (注 1)．また，近い将来，構築したコーパスを用いて，解説文生成シェアードタスクを開催することを計画している (注 2)．

（注 1）：http://nlp.ii.konan-u.ac.jp/fc_corpus.html
（注 2）：このため，現時点では，一部の結果のみ公開している．未公開の部分はシェアードタスク開催時に全て公開する計画である．

## 2. 関連研究

研究利用可能な学習者コーパスは年々増えている．これらの学習者コーパスが，学習者の言語データを対象にした言語処理（例えば，文法誤り検出／訂正）の発展に貢献していることは疑いようもない．

初期には，誤りの情報などが付与されていない生のコーパスデータの公開が主流であった．現在，そのような学習者コーパスは，ICLE [4]，NICT JLE [6]，ICNALE など多数利用可能である．これらの学習者コーパスは，学習者の言語に関する情報を与えてくれる重要な情報源である．例えば，単語の分散表現 [1] の獲得に利用されている．

2000 年ごろから，文法誤りや綴り誤りの情報が付与された学習者コーパスが公開され始めた．現在では，非常に多くの選択肢がある．例えば，NICT JLE（の一部），CLC FCE [15]，CoNLL-2013/CoNLL-2014 shared-task datasets [13], [14] などがある．関連したコーパスとして，訂正前後の文をペアにしたパラレルコーパスもある（例えば，Lang-8 learner corpus (Lang-8) [9]）．以上のコーパスのうち，Lang-8 と CoNLL-2014 が，それぞれ，共通の訓練データと評価データとして使用されることが多い．

別の方向性として，構文情報の付与がある．Nagata ら [10] は，学習者コーパス専用に句構造を定義し，句構造の付与を行っている．また，Berzak ら [2] は，同様に，依存構造付きコーパスを公開している．更に，両者とも構築したコーパスを用いて，学習者の英文向けの品詞解析器と構文解析器を開発している．

一方で，解説文の情報が付与された学習者コーパスで，一般に公開されたものは存在しない．その理由の一つは，解説文の付与には，多大な時間と労力を要するためであると想像される．研究に利用可能なデータが存在しないことと並行して，解説文自動生成の研究例も非常に少ない．一部の研究者 [7], [8] は，人手で作成した規則で文法誤りを診断することを試みている．また，Nagata ら [11] は，自動獲得した格フレームで，前置詞誤りの解説を記述する手法を提案している．これら過去の研究では，幅広い誤りに対して解説文を生成することは困難である．このような状況を改善し，解説文生成の研究を促進させるため，我々は解説文付きコーパスの開発に取り組んでいる．

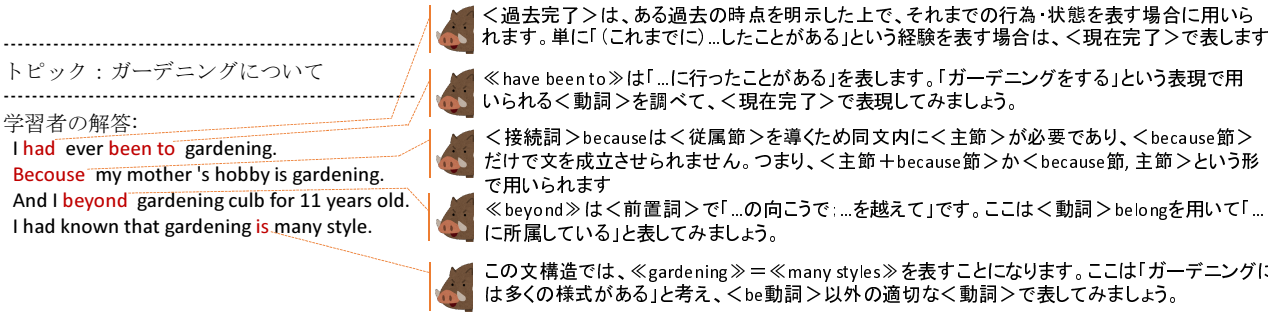


図 1： 一般解説文の例.

## 3. コーパスデザインとガイドライン

### 3.1 対象コーパス

1. で述べたように，基本となるコーパスは ICNALE written essay と KJ である．ICNALE は，解説文生成において好ましい特徴をもつ．特に，収録されているエッセイのトピックが統制されていることが重要となる．なぜなら，トピックの統制により，授業内のライティング課題や資格試験など，全ての学習者が同一のトピックについて書くという，語学学習で頻繁にみられる状況を想定できるからである．このような状況下で，どのような解説文が生成可能であるかを探求することは，技術的にも実用的にも有益である．ICNALE の具体的なトピックは，(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. の 2 種類である（両者とも argumentative essay である）．基本的に，この 2 種類のトピックについて，各学習者はライティングを行っている．更に，ICNALE は，その規模も比較的大きい (注 3)．書き手は，アジアの 10 の国と地域の大学生および大学院生である．能力レベルは，CERF の A2 から B2+と推定されている．

もう一方のコーパスである KJ は，その規模は小さい（233 エッセイ）が，他のコーパスにない特徴を有する．具体的には，綴り誤り，文法誤り，品詞，句構造の情報を収録する．これらの情報が，解説文自動生成に有益である可能性を考慮し，対象コーパスに含める (注 4)．

### 3.2 基 本 方 針

解説文としてどのような情報を付与するかということは，それほど自明ではない．選択肢は，文法誤り，文章構成，内容など多岐に渡る．また，誤りや不自然な表現に対する解説だけでなく，より良くするための助言や学習者の意欲を向上させるための励ましなども解説文に含めることも可能である．

どのような情報を付与すべきか（また付与可能か）ということを検討するため，予備的な付与作業を二度行った．各回，10 エッセイ，合計で 20 エッセイ，ICNALE と KJ からサンプリングし，二人の作業者が独立に解説文を付与した（その際，特に制限は設けず，自由に付与を行った）．作業者の一人は，英語の構文情報付与に 10 年以上従事しているプロのアノテータである．もう一人は，第一著者である．解説文の付与には，MS-Word のコメント機能を利用した．各回の終わりに，付与結果を検討するセッションを設け，ガイドラインのドラフトを策定した．

その結果，基本方針として，次の事項を定めた：

**基本方針：** 書き手のレベルにとって最重要と思われる項目に重点を置いて解説文を付与する．

この理由は次のとおりである．解説文の種類は多岐に渡り，網羅的に解説文を付与することは困難であることが予想される．また，書き手にとって，未習熟の項目や難しすぎる項目は，学習という観点から好ましくない．更に，学習者に，大量の解説文を一度に与えることも好ましくない．適切な難易度の解説を適量与えることが大切である．これらを考慮して，基本方針では，書き手のレベルにとって最も重要と推測される項目を中心にして，解説文付与を行うことを規定する．もちろん，解説として何が重要かということは，書き手のレベルやアノテータの主観による．そこで，我々の定めたガイドラインでは，アノテータの主観により，書き手のライティング能力を推定し，その能力に適していると思われる項目について解説文を付与するよう定めている．このため，構築されたコーパスは，文法誤り，語彙選択，文章構成，メカニクス（句読点の用法など）など多様な事項についての解説を含む（詳細は，4. で述べる）．なお，以降では，この解説文のことを一般解説文と表記する．

一般解説文に加えて，前置詞の用法に限定した解説文（以降，前置詞解説文と表記する）も付与することとした．これは，前置詞誤りは頻出することに加えて，その解説が比較的容易なためである．一般解説文と異なり，前置詞を対象にした場合，どこをどのように解説するかは，より明確である．

(注 3)：5,600 エッセイ，約 1,300,000 トークンである．各エッセイは，標準，200〜300 トークンから成る．

(注 4)：ただし，解説文として何が重要かということを一から検討するため，解説文付与の際には，それらの情報は利用しなかった．素のコーパスデータに対して，今回新たに解説文の付与作業を行った．

そのため，全ての前置詞誤りを付与対象とすることとした．加えて，アノテータが解説したい部分（例えば，より良くするための助言やモチベーション向上のための励まし）に対しても解説文を付与してよいこととした．なお，一般解説文と前置詞解説文は，独立に付与することとした．したがって，両者には前置詞に関する部分について一部重なりはあるものの，完全には一致しない．

付与に先立って，解説文の記述言語を決定しなければならない．基本的には，英語もしくは書き手の母語が候補となる．本研究では，次のことを考慮して日本語を第一記述言語とした：(1) 英文ライティングにおいて，別の言語（日本語）で解説文を生成することは，より技術要求が高く，興味深い；(2) 初級から中級の学習者は，英文ライティング学習中に，英語で解説を記述されると認知的負荷が高くなりすぎる可能性があり，母語で記述したほうが好ましい場合がある；(3) 全ての母語で解説文を記述することはコストが高いため書き手の母語のうち一つを選ぶ必要がある．ただし，データの利用範囲を広げるため，解説文には英訳を付けることとした．

### 3.3 付与手順とガイドライン

一般解説文の付与手順は次のとおりである：

（1） 解説文の付与に先立ち，文書全体を通読
（2） 書き手のレベルにおいて重要と思われる事項を決定
（3） （2）の結果に基づき，5〜10 個の解説文を付与 (注 5)
（4） 付与後，結果を再確認
（5） 必要に応じて結果を修正

前置詞解説文については，上記手順（2）を「前置詞誤りおよびその他解説が必要な箇所を特定」に置き換えて付与を実施する（全ての前置詞誤りを対象にしているため）．

予備的な付与作業の結果を基に，次の特殊タグを規定した．文法項目タグ（<, >）と引用タグ（<<, >>）である．前者は，タグ内の語句が文法項目であることを示す（例：< 動詞派生前置詞 >）．このタグにより，生成した解説文を文法書などの外部知識と紐付けることができる．後者は，解説文中の語句が，解説対象となっている英文から引用されていることを示す．解説文生成手法の開発において有益になることを期待して引用情報も含めた．

以上が，本コーパスのデザインとガイドラインの概要である．詳細は，コーパスデータに含まれるガイドラインを参照されたい．

## 4. コーパス構築

一般解説文の付与のために，12 名のアノテータを雇用した．アノテータは，英語教材の問題作成者，編集者，元編集者および英語資格試験の採点者のいずれかである．前置詞解説文では，別の 2 名を雇用した．こちらは，英語の構文情報などの付与の経験が 10 年以上あるプロのアノテータである．

雇用後，更に予備的な付与を 2 回実施し，正式版のガイドラインを策定した．ICNALE と KJ から 200 エッセイをサンプリングし，一般および前置詞解説文それぞれについて 2 名のアノテータが担当した．作業終了後，付与結果について議論を行いガイドラインを確定した．

以上を経て，正式な付与作業を開始した．一般解説文については，12 名のうちいずれかの 2 名が一つのエッセイに対して独立に解説文を付与した（したがって，各エッセイについて 2 バージョンの解説文データが存在する）．一方，前置詞解説文については，2 名のうちいずれかが一つのエッセイを担当した．なお，前処理として，Stanford Statistical Natural Language Parser (ver.2.0.3) [3] を用いて文分割とトークン分割を行った．

表 1 に，現在までに作業を終えたコーパスの統計量を示す．統計量の算出のため，MS-Word のコメント形式を TSV 形式（学習者の英文，解説文，オフセット（どの語句に対する解説文かを示すインデックス））に変換した（変換のためのツールも公開している）．なお，一般解説文の解説文数が二つあるのは，2 名のアノテータによる 2 バージョンの解説文が存在するためである．

付与結果の傾向を把握するために，人手により定性的および定量的な分析を行った．一般解説文の 2 バージョンの付与結果から，それぞれ 20 エッセイをサンプリングした．各バージョン，それぞれ 175 と 182 の解説文が付与されていた．対応するエッセイのペアについて，付与されている解説文が同じ内容かどうかを主観的に判断した．その結果，78 個について同じ内容と判断された．これは，Szymkiewicz-Simpson 係数 0.446 に相当する．この結果は，一般解説文付与作業の自由度は高いにもかかわらず，付与内容はある程度一致することを示す．言い換えれば，内容が大きく発散することはなく，自動生成できる可能性があることを示唆する．

前置詞解説文についても Szymkiewicz-Simpson 係数を求めたところ 0.678 とより高い値が得られた（2 名のアノテータが作業した全 200 エッセイについて，アノテーション箇所が一致するかどうかで求めた）．両者が完全に一致しないのは，単純な付与ミスに加えて，付与が任意となるケースが存在するからである（すなわち，より良くするための解説やモチベーションを向上させるための励ましなどである）．

更に詳細に分析するため，一般解説文を細分類した．新たに，169 の解説文をサンプリングし，解説内容に基づいて分類した．その結果を表 2 に示す．表 2 より，大部分の解説文は文法誤りに関するものであることがわかる．また，語彙選択（例：“tell” と “say” の使い分けの解説）が二番目に多く 16%を占めることがわかる．以降，文章構成 （例：「“On the

(注 5)：ICNALE に合わせて，対象文書は 200〜300 語から成ることを想定している．

表 1： 解説文付き学習者コーパスの統計量.

| 解説文 | 一般 | | 前置詞 | |
|---|---|---|---|---|
| コーパス | ICNALE | KJ | ICNALE | KJ |
| エッセイ数 | 752 | 233 | 1,092 | 233 |
| 文数 | 11,218 | 3,236 | 16,887 | 3,236 |
| トークン数 | 191,988 | 30,802 | 275,725 | 30,802 |
| 解説文数 | 6,410/6,413 | 2,005/2,037 | 3,034 | 538 |

表 2： 一般解説文の分類.

| 分類 | 割合 (%) |
|---|---|
| 文法誤り | 61.5 |
| 語彙選択 | 16.0 |
| 文章構成 | 10.7 |
| メカニクス | 10.1 |
| その他 | 1.8 |

other hand" は比較に用いる表現です．帰結には別の表現があるので辞書などで調べましょう。」），メカニクス（例：「複数の名詞句を列挙する際にはカンマが必要です。」）と続く．また，169 の解説文のうち 18（10.7%）については，褒める内容であった．例えば，複雑な構文構造の使用や論旨を明確にする接続詞の使用に対して褒める内容の解説文が付与されていた．

## 5. おわりに

本稿では，解説文生成技術の実現を目指して，構築した解説文データについて報告した．まず，基本方針とアノテーションガイドラインについて議論した．その議論に基づいて，実際に解説文付与を行った二つのコーパス（ICNALE と KJ）について詳細を述べた．構築結果の一部を公開しており，今後，解説文生成の研究に利用されることが期待される．

現在は，構築したデータを用いて解説文自動生成手法の開発に取り組んでいる．近い将来，解説文自動生成を題材としたシェアードタスクを開催する予定である．

## 謝　辞

本研究の一部は，JST，さきがけ，JPMJPR1758 の支援を受けたものである．

## 参考文献

[1] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," Proc. of 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.715–725, 2016.

[2] Y. Berzak, J. Kenney, C. Spadine, J.X. Wang, L. Lam, K.S. Mori, S. Garza, and B. Katz, "Universal dependencies for learner English," Proc. of 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.737–746, 2016.

[3] M.C. de Marneffe, B. MacCartney, and C.D. Manning, "Generating typed dependency parses from phrase structure parses," Proc. of 5th International Conference on Language Resources and Evaluation, pp.449–445, 2006.

[4] S. Granger, "The international corpus of learner English," in English language corpora: Design, analysis and exploitation, pp.57–69, Rodopi, 1993.

[5] S. Ishikawa, The ICNALE and Sophisticated Contrastive Interlanguage Analysis of Asian learners of English, pp.91–118, Kobe University, Kobe, 2013.

[6] E. Izumi, T. Saiga, T. Supnithi, K. Uchimoto, and H. Isahara, "The NICT JLE Corpus: Exploiting the language learners' speech database for research and education," International Journal of The Computer, the Internet and Management, vol.12, no.2, pp.119–125, 2004.

[7] J. Kakegawa, H. Kanda, E. Fujioka, M. Itami, and K. Itoh, "Diagnostic processing of Japanese for computer-assisted second language learning," Proc. of 38th Annual Meeting of the Association for Computational Linguistics, pp.537–546, 2000.

[8] K.F. McCoy, C.A. Pennington, and L.Z. Suri, "English error correction: A syntactic user model based on principled "mal-rule" scoring," Proc. of 5th International Conference on User Modeling, pp.69–66, 1996.

[9] T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto, "The effect of learner corpus size in grammatical error correction of ESL writings," Proc. of 24th International Conference on Computational Linguistics, pp.863–872, 2012.

[10] R. Nagata and K. Sakaguchi, "Phrase structure annotation and parsing for learner English," Proc. of 54th Annual Meeting of the Association for Computational Linguistics, pp.1837–1847, 2016.

[11] R. Nagata, M. Vilenius, and E. Whittaker, "Correcting preposition errors in learner English using error case frames and feedback messages," Proc. of 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.754–764, 2014.

[12] R. Nagata, E. Whittaker, and V. Sheinman, "Creating a manually error-tagged and shallow-parsed learner corpus," Proc. of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp.1210–1219, 2011.

[13] H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant, "The CoNLL-2014 shared task on grammatical error correction," Proc. 18th Conference on Computational Natural Language Learning: Shared Task, pp.1–14, 2014.

[14] H.T. Ng, S.M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, "The CoNLL-2013 shared task on grammatical error correction," Proc. 17th Conference on Computational Natural Language Learning: Shared Task, pp.1–12, 2013.

[15] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," Proc. of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp.180–189, 2011.