英語学習者コーパスのための句構造アノテーション

永田　亮†　坂口　慶祐††
† 甲南大学知能情報学部　†† Johns Hopkins University
E-mail: †nagata-nlp2015@hyogo-u.ac.jp, ††keisuke@cs.jhu.edu

1. はじめに

本稿では，英語学習者コーパスを対象にした句構造アノテーションについて述べる．本研究の目的は，学習者特有の言語現象に対して信頼性の高い句構造アノテーションを実現することにある．更に，本稿では，実際に句構造を付与した学習者コーパスについて考察を行い，提案手法の有効性と問題点について議論する．

学習者コーパス向けの構文アノテーションに関しては，既に多くの研究が存在する．Nagata ら [3] は，品詞／句情報をアノテーションする手法を提案している．Foster [2] は，構文情報付き母語話者コーパスに誤りを自動的に埋め込むことで，誤りを含むコーパスを生成する手法を提案している（ただし，学習者の傾向を真に反映したものではない）．また，Ragheb ら [4] は，学習者英語を対象とした依存構造アノテーション手法を提案し，依存構造アノテーションにおける問題点を明らかにしたという点で重要な位置づけにある．

一方で，学習者コーパス向けの構文アノテーション研究には未解決の問題も存在する．著者らが知る限り，学習者コーパスを対象にした句構造アノテーションは存在しない(注 1)．句構造は依存構造に比べて，(i) 語順に関する情報を直接取り扱える，(ii) 構文構造を句対句という抽象的なレベルで表現できるというメリットがある．(i) の性質により，語順の誤りを含む学習者の英文に対し，どのような句で語順の誤りが起こりやすいかを明らかにすることができる．更に，(ii) により，学習者の特徴を句構造規則や構文木で表現できる．実際に， 5. で示すように，我々が構築したコーパスにおいて学習者特有の句構造規則と木構造を発見することができた．このように，学習者の特徴を句構造規則，構文木として分析することは第二言語習得の観点からも興味深い．本研究の最終的な目的は，母語話者の句構造規則に，どのような規則を追加／削除することで学習者の英文を生成することが可能となるのかを明らかにすることである．

また，学習者コーパスを対象とした依存構造／句構造アノテーションに共通した問題点として，コーパスが一般に公開されていないという問題もある．誤り情報が付与されたコーパスの公開により文法誤り検出の性能が飛躍的に向上したように，構文情報付き学習者コーパスの公開は関連する分野に大きく貢献するであろう．

そこで，本稿では，学習者英語を対象とした句構造アノテーション手法を提案する．本研究の貢献は次の 3 点である：(1) 学習者特有の言語現象に対して一貫性の高いアノテーションを実現するための手法を提案する；(2) アノテーション精度と学習者特有の句構造規則を示すことにより提案手法の有効性を検証する；(3) 構築したコーパスを公開する(注 2)．

2. 基本方針

アノテーション規則を策定するに先立ち，基本方針として次の 5 項目を制定した：

(P1) 一貫性の重視
(P2) 規則集合の最小化
(P3) 表層形の重視
(P4) 編集距離の最小化
(P5) 直感に基づいたアノテーション

「(P1) 一貫性の重視」は，提案アノテーション手法では，一貫性を最も重視するということを述べたものである．一般に，アノテーション規則の複雑さと得られる情報量はトレードオフの関係にある．複雑な規則は，より詳細な情報を提供するかもしれないが，一貫性の高いアノテーションが行えなければ，得られるコーパスの価値は低い．そこで，アノテーション規則を作成する際に複数の候補がある場合は，一貫性が高くなるものを採択することとした．

「(P2) 規則集合の最小化」も，アノテーションの一貫性に関するものである．句構造規則の数が増加するにつれて，アノテーション作業は複雑になる傾向にある．そこで，新たな句構造規則を追加する際には，規則の数が少なくなる方法を採択する方針とした．ここで，この方針は規則集合全体に及ぶことに注意する必要がある．新しい句構造規則の追加は，既存の規則の修正，更なる規則の追加を引き起こす場合があるが，これらの全てを考慮して，規則集合の変化が極力少なくなる方法を採択することとした．

「(P3) 表層形の重視」は，トークンの表層形に従いアノテーションを行うことを要求する．学習者コーパスでは，誤りに起因して一つのトークンに対して複数のアノテーション候補が考えられることがある．例えば，“*My university life

(注 1)：Foster [2] の手法では，句構造付きのコーパスを生成できるが，上述の通り学習者の傾向を真に反映したものではない．

(注 2)：詳しくは，http://nlp.ii.konan-u.ac.jp/ を参照のこと．

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

is enjoy.” では，“enjoy” は表層形に従うと動詞，文脈に従うと形容詞や名詞などが候補となる．このとき，(P3) は表層の品詞である動詞を選択することを要求する．これは，上の例のように，文脈に従うと選択肢が複数となることが多く一貫性を低下させる原因となるためである．

「(P4) 編集距離の最小化」は，誤りを含む文から正しい文を復元する方法を規定する．アノテーションに際して，正しい文が必要となることは少なくないが，通常，複数の訂正候補が考えられる．(P4) は，訂正前後の編集距離が最小となるものを正しい英文として選択することを要求する．

以上の基本方針が適用できないケースについては，「(P5) 直感に基づいたアノテーション」を行うこととした．ただし，一貫性の低下を避けるため，基本方針の優先順位は (P1) – (P5) の順とする．

3. アノテーション規則

本アノテーション規則で基礎となるのは，学習者の英文を対象とした品詞／句情報アノテーションガイドライン [3] である．このガイドラインでは，Penn Treebank II-style bracketing guidelines [1]（以降，PTB-II と省略）のタグセットを採択している．従って，提案手法も，同じタグセットとアノテーションガイドラインに準じる．ただし，PTB-II の機能タグ，副詞タグ，空要素については対象外とする．

これらのタグセットとアノテーションガイドラインをもとに，学習者英語向けのアノテーション規則を作成するにあたり，大きな問題となるのは文法誤りである．文法誤りは，脱落，余剰，置換の 3 タイプに分類されることが多く，本稿でも，この 3 タイプに分けてアノテーション規則を説明する(注 3)．なお，規則に関する完全な記述については，コーパスに付随したアノテーションガイドラインを参照されたい．

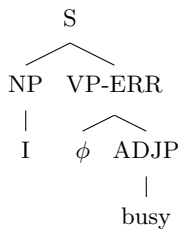
3.1 脱落タイプのアノテーション

脱落タイプとは必要な語句が脱落している誤りのことである．例えば，“*I am student.” では限定詞の脱落がある．

幸い，既存の PTB-II により大部分の脱落タイプの誤りを取り扱うことができる．例えば，上述の例では，限定詞の脱落に関係なく，表層形に従い次のようにタグ付けが行える：(S (NP I) (VP am (NP student)).)(注 4)．

一方，句の主辞（head）の脱落は問題となることがある．例えば，“*I busy.” では，動詞が脱落しているため，文を表す S 構造をアノテートすることができない．

この問題を解決する手段として，機能タグ “-ERR” を提案する．この機能タグは，当該句の主辞が脱落していることを表す．具体的には，対応する句のタグに “-ERR” を付け，主辞の脱落を “ϕ” により表す．上述の例の場合，



となる．この機能タグにより主辞が脱落していたとしても，既存の PTB-II をそのまま適用することができる．

主辞の脱落を同定するためには正しい英文を復元する必要がある．その際に，基本方針（P4）が重要な役割を果たす．例えば，“*I want to happy.” では，訂正候補として少なくとも “I want to be happy.”（編集距離 1）と “I want happiness.”（同 3）の二つを想定することができるが，(P4) により，前者を採択し，脱落があるとする：“(S (NP I) (VP want (VP (PP to) (VP-ERR ϕ (ADJP happy)).)”．

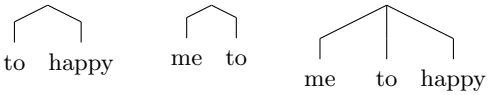
3.2 余剰タイプのアノテーション

余剰タイプとは，不要な語が挿入された誤りのことである．例えば，“*She discussed about it.” では下線部が不要である．

この例において，“about” の品詞はそれほど自明でない．表層形からは前置詞と解釈することができる一方で，“discuss” は前置詞を必要としないことから前置詞でないと主張することも可能である．この例のように，余剰タイプの誤りは，表層形と文脈が要求する品詞に不一致を引き起こす．

そこで，基本方針（P3）を利用して，表層形に従い品詞を決定する．よって，上述の “about” は前置詞とする．より上層の構造は，PTB-II を表層形の品詞に適用して決定できることが多い（このことは，同時に（P2）も満たす）．上述の例の場合，“(S (NP She) (VP discussed (PP (IN about) (NP it)))).)” とタグ付けできる．

一方で，既存の PTB-II を適用できない場合もある．具体例として，“*It makes me to happy” を考えてみる．(P3) により，“*It/PRP makes/VBZ me/PRP to/TO happy/JJ” と容易に品詞を決定できる．しかしながら，PTB-II には品詞列 “TO JJ” に適用できる規則がなく，新たな句構造規則が必要となる．このとき，三つの単語をまとめて句とする方法には次の三種類がある：



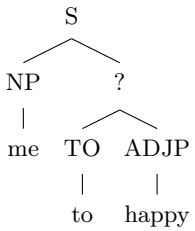
この中で，最初の案が最も直感にあう．より形式的には，次のように説明することができる．二番目の案は後置詞を暗に仮定する．これは，英語の前置詞システムと競合し，PTB-II の規則の広範囲に影響が及ぶ（(P1) に反する）．同様に，三番目の案は，代名詞，前置詞，形容詞から新たな句を生成することに相当し，PTB-II への影響が大きい．一方で，一番目の案では，前置詞は別の句（ただし，通常は形容詞句とでは

(注 3)：実際には，語順の誤り，フラグメント，メカニクスの誤りなどに対応するためのアノテーション規則も存在する．

(注 4)：これ以降の例では，議論に直接関係ない部分の構造は略記する．

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

ないが）と前置詞句を生成できることを考慮すると PTB-II への影響が最も少ない．また，一番目の案では，



と示されるように，三つの単語から S を生成できるため，より上層の構造との整合性が高い．よって，一番目の案を採択することとする．残された問題は，当該句の名前を決定することである．候補として，上層の S を導ける前置詞句と形容詞句が考えられる．前述の通り，前置詞は別の句と共に前置詞句を構成できることを考慮すると前置詞句とするのが妥当そうである．また，句の主辞性を考慮すると，直接の子要素に主辞になりえる語がある場合，その語の品詞で句の名前を決定することは自然である．以上の議論により，当該句を前置詞句とする．最終的に，上述の例文は，“(S (NP It) (VP makes (S (NP me) (PP (TO to) (ADJP happy)))))” とタグ付けする．

これらの手順は，次のように一般化できる：(a) 直感的にまとまりが良い句を採択する（ただし，その際には新たに追加される規則の数を最小とする方法を採択），(b) 句のラベルは，その句内の直下の語（語がない場合は句）のうち主辞となることができるもののタグにより決定する．

3.3 置換タイプのアノテーション

置換タイプとは，別の語に置き換える必要がある誤りを指す．置換後も同じ品詞となる誤りであれば，既存の PTB-II がそのまま利用できる．従って，新たな句構造規則を作成する必要はない．

一方，二つの品詞に渡る置換の場合，表層形と文脈が要求する品詞に不一致が生じるため特別な処置が必要となる(注 5)．例えば，“*I went to the see.”（正しくは “sea”）では，動詞と名詞の間で曖昧性がある．提案手法では，従来研究 [3], [4] でも採択されている二層のタグ付けを行う．二層のタグ付けとは，表層形／文脈の品詞両方をタグ付けすることである．提案手法では，特別なタグ “CE：表層の品詞：文脈の品詞” で 2 種類の品詞をタグ付けする(注 6)．CE より上層の構造については，文脈の品詞に基づいて決定する．

基本的には，3.1– 3.3 で述べた方法により句構造アノテーションを行うが，これらの規則が適用できない言語現象も存在する．その場合，UK と XP タグで対応する．UK は表記の誤りなどにより品詞が特定できない語に使用する．XP は，誤りにより句の境界や主辞が不明である際に使用する．

4. コーパスアノテーション

提案手法を利用して，句構造アノテーションを行った．対象コーパスとして，Konan-JIEM learner corpus [3]（以降，KJ コーパスと省略）を選択した．KJ コーパスには，品詞と句の情報が人手で付与されており，本作業では，その情報をそのまま利用した．更に，主辞の脱落のアノテーションの際には，同コーパスの文法誤りの情報も利用した．ただし，アノテーションの効率を考慮し，本作業では主動詞と前置詞の脱落のみアノテーションした．これは，その他の主辞脱落に対しては，既存の PTB-II が適用できる場合が多いためである．適用できない場合は，XP で対応することとした．

アノテーション作業には，プロのアノテータ二人が参加した．まず，一人目の作業者が全データのアノテーションを行った．作業終了後，結果に関して作業者と著者で議論を行い，アノテーション規則を一部修正した．修正後の規則を用いて，同じ作業者がアノテーション結果の再チェックを行った．次に，コーパスから 30 文書をサンプリングし，トライアルセット（11 文書，122 文，919 トークン）と評価セット（19 文書，211 文，1,785 トークン）に分割した．最後に，二人目の作業者が両セットのアノテーションを行った．トライアルセットについては，必要に応じて最初のアノテータと相談することを許し，作業終了時に，二人の結果の差異を提示した．評価セットは独立に作業を行った．

全ての作業終了後，二人の作業者間の一致率を求めた．具体的には，一人目の作業者の結果を正解データとみなして，再現率，適合率，F 値，完全一致率を求めた（EVALB(注 7) とパラメタ COLLINS.prm を用いた）．

表 1 に，評価結果を示す．表 1 より，一致率が非常に高いことがわかる．このことは，誤りが頻出する KJ コーパス（平均誤り率 1.3 個／文）に対して，提案手法で一貫した句構造アノテーションを実現できたことを示す．実際，文法誤りに起因したアノテーションミスは非常に少ない結果となった．不一致の最も大きな原因は，主語と主動詞の間にある副詞句の解釈であった（例：I often go）．PTB-II では，動詞句の子要素，S の子要素どちらとしてもよいため，二人の作業者間で不一致が頻繁に見られた（従って，厳密にはアノテーションミスとは言えない）．一致率が非常に高い別の理由として，KJ コーパスでは，平均文長が比較的短く（平均 8.0 語），複雑な構文が少ないことも挙げられることに注意する必要がある．書き手のレベルが向上し，構文が複雑になったとしても同様な一致率が達成できるかは更なる調査が必要である．今回の結果は，少なくとも KJ コーパスと同レベル（novice〜

(注 5)：正確には，NN と NNS のような置換であれば，PTB-II が適用可能である．問題となるのは，名詞と動詞のように品詞の大分類が異なる場合である．詳しくは，ガイドラインを参照のこと．

(注 6)：実際のアノテーションでは，XML を用いて二種類の品詞をコーディングする（例：<CE suf=”VB” con=”NN”>see</CE>）．

(注 7)：http://nlp.cs.nyu.edu/evalb/

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

表 1： アノテーションの一致率.

Set	再現率	適合率	F 値	完全一致率
Trial	0.980	0.981	0.980	0.910
Test	0.966	0.980	0.973	0.853

intermediate）の英文に対しては一貫性の高いアノテーションが実現できることを示唆する.

5. 考　　察

前節で，一貫性の高い句構造アノテーション結果が得られたが，このアノテーションが学習者の特徴を適切に記述しているかは更なる議論が必要である．一貫性を重視する提案手法ではシンプルな規則を好むため，学習者の特徴を十分に反映していない可能性もある.

そこで，アノテーション結果から特徴的な句構造規則を抽出するというパイロットスタディを行った．抽出の基本アイデアは，KJ コーパスと母語話者コーパス（Penn Treebank-II），それぞれから得られる句構造規則を比較するというものである．句構造規則を $A \to B$，その条件付き確率を $p(B|A)$ としたとき， $A \to B$ の特徴度を

$$s(A \to B) = \log \frac{p_L(B|A)}{p_N(B|A)} \qquad (1)$$

と定義した（L と N により学習者コーパスと母語話者コーパスを区別する）．条件付き確率は，予期尤度推定法により推定した．頻度を求める際には，機能タグ，副詞タグ，空要素を削除し，コーパス間の差異を極力減らすようにした．また，統語的特徴をより明確にするため，終端記号及び前終端記号のみから成る規則は対象外とした．更に，今回用いた母語話者コーパスに頻出する数量表現を含む規則（例：(NP (QP 100 million) yen)）も対象外とした.

表 2 に，特徴度の昇順／降順でソートした句構造規則の上位 10 件を示す．昇順／降順のカラムは，KJ コーパスで過剰／過小使用される規則に対応する.

過剰使用カラムでは，主辞の脱落を示す ϕ が多く含まれることがわかる．このことは，母語話者コーパスでは ϕ が定義されないことを考慮すると驚くべきことではない．しかしながら，表 2 に示される規則は，どのような統語環境で主辞の脱落が起こりやすいかも明らかにする．例えば，名詞句の前に位置する前置詞の脱落に加え，S を導く前置詞句における脱落も明らかにしている（例：*I am good _ playing football.）．より興味深い規則として “VP $\to \phi$ ADJP” を挙げることができる．実際にどのような文でこの規則が使用されているかコーパスを調査したところ， 3.1 で構文木を示した “*I busy.” のような形容詞述語文における動詞の脱落に対応することが明らかとなった．このことは，「形容詞述語文ではコピュラが省略可能である」という日本語の性質が母語干渉として英語に転移したと分析できる.

表 2： 特徴的な句構造規則.

過剰使用	過小使用
PP $\to \phi$ NP	S → S , NP VP .
S → XP VP .	S → S : S .
VP $\to \phi$ ADJP	NP → NP JJ NN
PP $\to \phi$ S	S → ADVP VP
PP → TO UP	VP → VB VP
SBAR → IN NN TO S	NP → NP NP
S → XP .	NP → NP , NP
S → ADVP NP ADVP VP .	VP → VBD SBAR
VP → VBP S ADVP	VP → ADVP VBN PP
VP $\to \phi$ NP	VP → VBN S

一方，過少使用のカラムは，KJ コーパスでは複雑な構文の使用が少ない傾向にあることを示す．例えば，右辺に S, SBAR，VBN を含む規則が半数を占めるが，いずれも複雑な構文を導入する規則である．KJ コーパスの書き手は，これらの規則を十分に習得していない可能性が高い．従って，これらの規則を（学習者が理解しやすい形で）提示することは，効果的な学習支援となると期待できる．このように，句構造アノテーションは，学習者英語の特徴を明らかにするだけではなく，学習支援のための重要なデータにもなる.

以上の通り，シンプルな手法に基づくにもかかわらず，得られた結果は示唆に富むものである．このことは提案手法で学習者の特徴（少なくともその一部）を適切に記述できることを示す．同様な分析を構文情報付きコーパスなしで網羅的に行うことは困難であろう.

6. おわりに

本稿では，英語学習者コーパス向けの句構造アノテーションについて述べた．本研究では，(1) 学習者特有の言語現象に対して一貫性の高いアノテーションを実現するための基本方針とアノテーション規則を提案し，(2) アノテーション精度と学習者特有の句構造規則を示すことにより提案手法の有効性を示した．現在，作成したコーパスを公開中である．今後は，より洗練された抽出手法（例えば，一部語彙化した規則や部分木のマイニング）を用いて，更なる学習者の特徴解明に取り組む予定である.

参考文献

[1] A. Bies et al., “Bracketing guidelines for Treebank II-style Penn Treebank project,” 1995.

[2] J. Foster, “Treebanks gone bad: generating a Treebank of ungrammatical English,” 2007 Workshop on Analytics for Noisy Unstructured Data, pp.39–46, 2007.

[3] R. Nagata et al., “Creating a manually error-tagged and shallow-parsed learner corpus,” Proc. of 49th ACL, pp.1210–1219, 2011.

[4] M. Ragheb et al., “Defining syntax for learner language annotation,” Proc. of 24th COLING, pp.965–974, 2012.

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.