英語論文の執筆を支援する定型表現集のカテゴリ構造の分析

岩月 憲一† 相澤 彰子‡†
†東京大学大学院情報理工学系研究科
‡国立情報学研究所
{iwatsuki, aizawa}@nii.ac.jp

1 はじめに

非英語母語話者が英語論文を執筆する際には，使用する表現が適切であるかを調べるのに相当な時間を費やす。学術論文中で繰り返し使われる定型表現（formulaic language）は，このような執筆時の負担を軽減するために有用であるとされ，English for Specific Purposes (ESP) の分野では，定型表現を論文中から抽出することが行われてきた。Simpson-vlach らは，論文コーパス中から n-gram を抽出し，その中から有用な表現を人手で選択して定型表現リストを作成している [10]。Vincent，Gray らは，定型表現中に，任意の単語を入れることができるスロットを含めた phrase frame を抽出している [9,11]。さらに Brooke らは，スロット幅を 4 語まで拡大し，品詞タグを用いた正規表現によって表現する手法を提案している [13]。また，ウェブ上で簡単に参照できる定型表現集がいくつか存在する [1–3]。これらの表現集は，定型表現をいくつかのカテゴリに分類している。以下に例を挙げる。

「論文の目的」カテゴリ

- *The purpose of this study is ...*
- *This study provides ...*

しかしながら現実には，定型表現集を活用して論文を執筆するのは，必ずしも容易ではない。ユーザは定型表現を定型表現集中から検索する際に，定型表現そのものを検索クエリにすることはできない。そのため，表現集のカテゴリを頼りに使えそうな定型表現を検索することになるが，表現集によって異なるカテゴリ構造を覚えるのは困難であり，ユーザの意図に合致するカテゴリを見つけるのに手間がかかる。

定型表現のカテゴリには，2 種類考えられる。1 つは，定型表現そのものが持つ機能に基づいて分類する手法である。例えば，Biber ら [12] は，定型表現を stance expressions，discourse organizers，referential expressions の機能に分け，さらにそれぞれのカテゴリの中にサブカテゴリを設けている。もう 1 つは，論文全体の構造に基づいて分類する手法である。Swales [4,5] は，論文の各セクションが複数の move から構成されており，各 move はいくつかの step からなると分析している。本稿での要件，すなわち執筆時にユーザが定型表現を検索するという観点からは，後者のカテゴリ構造の方が適していると考えられる。

以上の検討を踏まえ，本研究では，既存の定型表現集のカテゴリ構造と move に基づいた構造を分析し，ユーザにとって使いやすいカテゴリ構造を検討する。

2 アプローチ

既存の定型表現集のカテゴリ構造は様々である。階層化されているカテゴリ構造を持つ場合，最も深い階層にあるカテゴリの名称を，定型表現と結びつける。

$$c_i \rightarrow \{f_0, f_1, ...\} \tag{1}$$

ここで，c_i は最下層のカテゴリであり，f_n は定型表現である。先の例の場合，c_0 = 論文の目的，f_0 = *The purpose of this study is ...*，f_1 = *This study provides ...* となる。

続いて，定型表現を論文の構造に基づいて分類することを考える。ここでは，論文の各セクション（abstract を含む）が持つ move，step [4,5] と，定型表現を結びつけることを考える。

$$s_i \rightarrow \{f_0, f_1, ...\} \tag{2}$$

ここで，s_i はある move 中の step であり，f_n は定型表現である。各 step は，1 つ以上の定型表現と対応している。

本稿では，s_i と c_i の関係について分析を行う。既存の定型表現集の各カテゴリに含まれる定型表現が，各 move にどのような分布で存在するかを調べる。

3 実験

3.1 既存の定型表現集の分析

ウェブ上で利用可能な既存の定型表現集を 3 種類用いた [1–3]。それぞれの特徴は表 1 の通りである。「Sentence Starters, transitional and other useful words」（表現集 1）[1] は *This essay discusses ...* や *In summary, ...* など，短い定型表現が多い。逆に，「Phrasebank」（表現集 2）[2] は *One of the limitations with this explanation is that it does not explain why...* のような長い表現が多い。「『英語論文に使う表現文例集』のレジュメ」（表現集 3）[3] は両者の中間的な長さである。

表 1: 既存の定型表現集の特徴

	表現集 1 [1]	表現集 2 [2]	表現集 3 [3]
カテゴリ数	11	134	39
項目数	213	1,587	233
平均単語数	2.18	10.5	6.25

ここで，既存の定型表現集に含まれる定型表現には，以下のような制約がある。そもそも収録されている定型表現は，ユーザが文脈に合わせて変形して使用することを想定しているため，実際のコーパス中でその定型表現が使われているかを，単純な文字列一致で調べることはできない。

まず，既存の定型表現集に含まれる定型表現には，例えば *Mike* や *discussions in legal and moral philosophy* などの単語（列）が含まれているが，これらは特定の文脈に特化しすぎており，コーパスを検索してもほとんどヒットしないし，執筆支援の上でも不要な表現である。こうした表現を取り除く必要がある。

次に，定型表現中に含まれる動詞について，これを表層形で用いるべきか，lemmatize 等の処理をして用いるべきかは検討すべき事項である。例えば，論文中では base という動詞はほとんどが based の形で用いられている。他方で，例えば*... is shown ...* という定型表現があるが，これを show, showed, showing と同一視すると，全く機能の異なる英文がヒットする可能性がある。ここでは，動詞の表層形を用いることとする。

そして，定型表現には，任意の語句を入れることができる空欄部分を持つものがある。例えば，*Before we show ..., it will be useful to discuss ...* や，A *primary concern of X is ...* などの，... や X の部分である。これについては，正規表現で書き換えることとする。

続いて，既存の定型表現集のカテゴリ構造について述べる。「Sentence Starters, transitional and other useful words」（表現集 1）[1] のカテゴリ構造は，*To introduce* や *To conclude* のように，定型表現がもつ機能によるものである。「Phrasebank」（表現集 2）[2] のカテゴリ構造は，*introducing work* や *describing methods* 等，論文のセクション構成に従っているものと，*being crucial* や *compare and contrast* 等，定型表現の機能に従っているものが混合している。「『英語論文に使う表現文例集』のレジュメ」（表現集 3）[3] のカテゴリ構造は，まず「前文」「先行研究」などの論文のセクション構成に基づくカテゴリがあり，それぞれの中に，細分化したカテゴリがある。例えば，「先行研究」カテゴリには，「先行研究レビュー」，「先行研究を評価」などのサブカテゴリがある。

以上の分析を踏まえ，既存の定型表現集の定型表現に対する前処理を行った。まず，人手で人名を除去し，任意の文字列を表す正規表現に置換した。次に，先頭と末尾の... と句読点は除去した。そして，A，B，X などの表現は任意の文字列を表す正規表現に置換した。さらに，数値は任意の数値を表す正規表現に置換した。最後に，study/paper などのように，複数の単語の候補が 1 つの定型表現に含まれる場合，それぞれが適合するように調整した。

前処理後の各表現集の定型表現が，データセットの各 move に含まれる数を表 2 に示す。

表 2: データセット中に含まれる既存表現集の定型表現

	background	method	result	conclusion
表現集 1	14,631	7,858	25,575	12,779
表現集 2	316	230	531	515
表現集 3	5,587	4,128	1,691	6,596

3.2 データセット

論文の move，step の構造に関しては，論文全体を対象とした研究 [8] や，abstract を対象とした研究 [6,7] がある。しかし，論文全体を対象として move，step のタグ付けが為されたコーパスは存在しない。そこで，本研究では，abstract を対象として分析を行う。

abstract 中の move，step の構成に関しては以下のような例がある。

- IMRD structure [6]

Move1 Introduction

Move2 Methods

Move3 Results

Move4 Discussion

- [7]

Move1 abstract

Step1 background of research

Step2 purpose

Step3 methods

Step4 results

Step5 conclusion

本稿では，PubMed[1] の structured abstract を用いる。Structured abstract は，予め論文の著者が background 等のラベルを段落に付与した論文抄録である。ここでは，最も一般的な，background，method，result，conclusion の構造を持つ抄録を採用した [14]。これを文単位で分割し，データセットとする（表 3）。

ここでは，background，method，result，conclusion の 4 つを move とし，これらと既存の定型表現集のカテゴリを結びつける。なお，abstract は論文本文と比べて短いため，ここでは各 move に対応する step は 1 つであるとする。

表 3: 使用したデータセットについて

	background	method	result	conclusion
文の数	49,526	59,257	86,767	39,183

4 実験結果

既存の定型表現集のカテゴリに含まれる定型表現が，データセットの各 move にどの程度含まれるかを数えた（図 1, 2, 3）。図 1，3 については，定型表現が 1 つも含まれなかったカテゴリを省略している。

まず，表現集 1 では，どれか 1 つの move に偏っているカテゴリを見いだすのは難しい。他方で，*To present inconclusive ideas* や *To present uncommon or rare ideas* は，method ではほとんど使われないことが分かる。

次に，表現集 2 では，多くのカテゴリが特定の move に偏っているように見える。ただ，表 2 に見るように，そもそもヒットした英文が少ないことが影響していると考えられる。

最後に，表現集 3 では，「結論」と「先行研究で分かっていない」カテゴリが，ほぼ 1 つの move と対応

[1] https://www.ncbi.nlm.nih.gov/pubmed

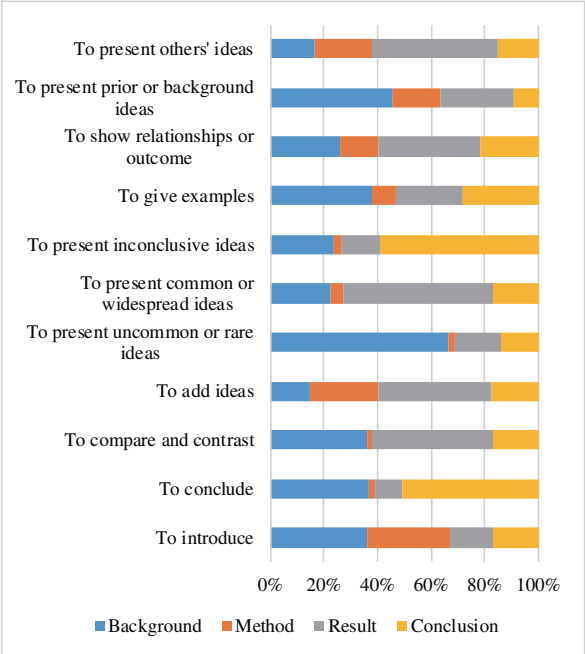

図 1: 既存の定型表現表現集 [1] のカテゴリと move の対応

している。他方，「注意喚起」や「定義」カテゴリは，4 つの move に満遍なく分布している。

5 おわりに

以上の分析から，既存の定型表現集のカテゴリ構造には，move と対応が取れるカテゴリと取れないカテゴリが存在していることが示唆される。この構造では，ユーザは執筆時に複数のカテゴリをまたいで定型表現を探す場合が少なくないと考えられる。より使いやすいカテゴリ構造を持った表現集が必要であり，今後は move 構造に沿った定型表現集の構築に取り組む予定である。

参考文献

[1] Sentence Starters, transitional and other useful words. http://www2.eit.ac.nz/library/ls_guides_sentencestarters.html.

[2] Morley, J. Academic Phrasebank. http://www.phrasebank.manchester.ac.uk.

[3] Kamegaya, M. (1997). 「英語論文に使う表現文例集」のレジュメ. http://kame.la.coocan.jp/ei_index.htm.

[4] Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

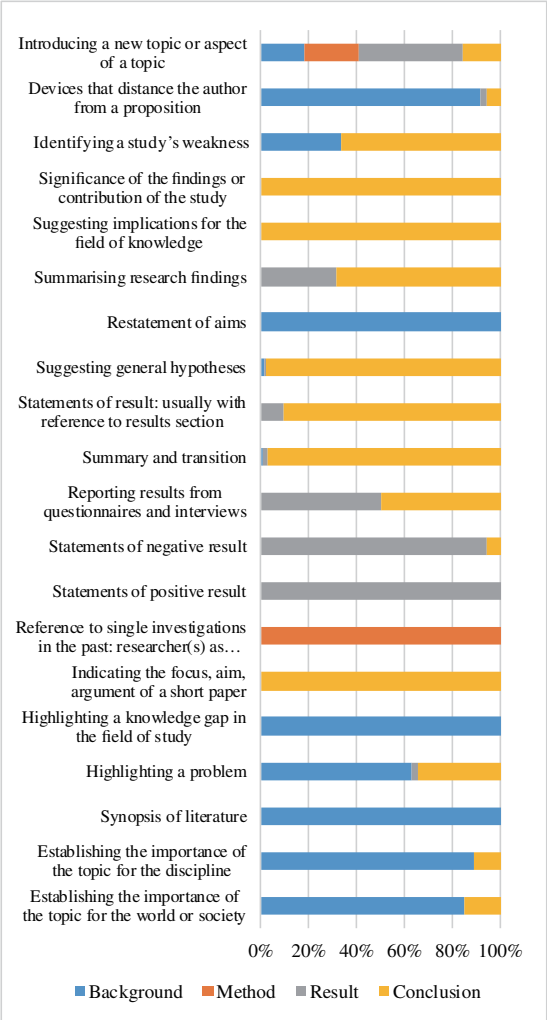

図 2: 既存の定型表現表現集 [2] のカテゴリと move の対応

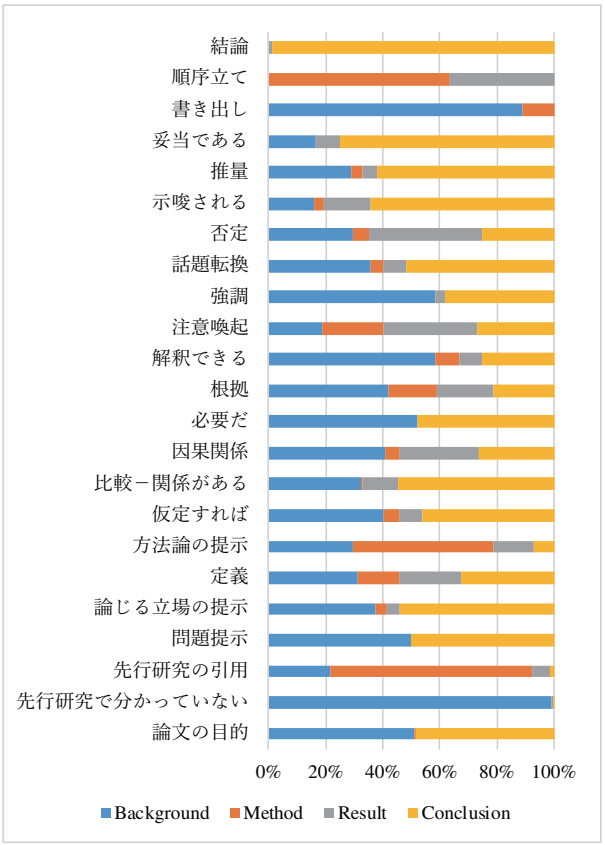

図 3: 既存の定型表現表現集 [3] のカテゴリと move の対応

[5] Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

[6] Lores, R. (2004). On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes. 23*. 280–302.

[7] Maswana, S., Kanamaru, T., & Tajino, A. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. *Ampersand. 2*. 1–11.

[8] Cotos, E., Huffman, S., & Link, S. (2015). Furthering and applying move/step constructs: Technology—driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes. 19*. 52–72.

[9] Vincent, B. (2013). Investigating academic phraseology through combinations of very frequent words: A methodological exploration. *Journal of English for Academic Purposes. 12*. 44–56.

[10] Simpson-vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics. 31*(4). 487–512.

[11] Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics. 18*(1). 109–135.

[12] Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics. 25*(3). 371–405.

[13] Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G., & Shein, F. (2015). Building a Lexicon of Formulaic Language for Language Learners. In *Proceedings of the 11th Workshop on Multiword Expressions*. 96–104.

[14] Willot, P., Hattori, K., & Aizawa, A. (2015). Extracting Structure from Scientific Abstracts Using Neural Networks. In R. B. Allen, J. Hunter, & M. L. Zeng (Eds.), *Digital Libraries: Providing Quality Information* (pp. 329–330). Springer.