

英語コーパス学会 大会予稿集2022

**PROCEEDINGS
OF THE JAECS
CONFERENCE 2022**

2022.10.1

ISSN 2436-6447

JAECS
Japan Association for English Corpus Studies

目次

COOPER Christopher Robert The Pre-processing of YouTube Transcripts for Corpus-based Spoken Language Analysis	... 1
仁科 恭徳／赤瀬川 史朗 日英・英日パラレルコーパス検索ツール『パラレルリンク』(Ver.1.20) －インターフェース, 検索機能, 活用研究などについて－	... 7
西村 知修 受動文における様態副詞の生起位置に関する一考察	... 13
藏菌 和也 起動表現の意味と補部にくる語との関係性に関する考察 －start NP を例に－	... 19
澤口 遼 日本人中学生英語学習者の take の使用と語義指導への提言	... 25
山口 一華／投野 由紀夫 学習者コーパスにおける品詞連鎖の分析による母語の影響に関する研究 －名詞句構造に焦点を当てて－	... 31
浅野 元子／松田 紀子 研究論文における英語と日本語のアブストラクトは読み手を意識しているか	... 37
USAMI Hiroko Development of the Ballet English Corpus (Ver. 1.0) for Art Major Students	... 43
田畑 圭介 COCA における正誤表現と非標準表現について	... 49
福本 広光 米国一般教書演説に出現する分離不定詞の効果に関する予備的研究	... 55

島津 美和子	
憲法をめぐる日本の帝国議会・国会会議録における「翻訳」の用法および 米国議会会議録における translation の用法の分析	... 61
藤原 康弘／岩男 考哲／伊藤 創／仲 潔	
日本人英語学習者の「うなぎ文」の使用に関する分析	... 67
佐竹 由帆	
英作文における動詞-名詞コロケーション産出に対する DDL の効果	... 73
FUKUDA Kohei／TONO Yukio	
A Critical Evaluation of the Optimal Association Measures for Creating L2 Learners' Collocations Lists	... 79
梶山 達也	
教科書コーパス分析における推定周辺平均値の有用性について ー英語法助動詞の頻度と意味の分析からー	... 85
森下 裕三	
分布意味論の手法を応用したフレーム意味論の分析	... 91
IKEO Reiko	
Negated Speech and Thought Presentation in Contemporary Present-tense Fiction	... 95

The pre-processing of YouTube transcripts for corpus-based spoken language analysis

COOPER Christopher Robert (Rikkyo University)

cooper@rikkyo.ac.jp

Abstract

The pre-processing steps taken to prepare YouTube transcripts for a multi-dimensional analysis are described. Examining the texts and keeping the goal and purpose of the study in mind is essential when deciding what elements of texts should be edited. The specific characteristics that needed to be cleaned are outlined, solutions to the problems are addressed, and the accuracy and suitability of the solutions is reported.

Keywords

pre-processing, spoken language, YouTube, tagging accuracy

1. Introduction

Recently YouTube is extremely popular and has potential for language learners to increase the amount of time they interact with the spoken form of the target language. The problem lies in identifying suitable videos. This paper is a preliminary report of extracting YouTube videos for EFL learners based on criteria such as register analysis as well as vocabulary demands. The process of preparing YouTube transcripts to be tagged with the MAT Tagger (Nini, 2019) for a multi-dimensional analysis is described. With the MAT Tagger, basic tags are assigned with the Stanford Tagger, then more fine-grained tags are added based on the linguistic features used in Biber (1988). The corpus design and multi-dimensional analysis will be described in more detail in future papers. However, to give some background, a near random sample of YouTube transcripts was downloaded using Python with the YouTube Data API (<https://developers.google.com/youtube/v3>) and the YouTube Transcript API (<https://pypi.org/project/youtube-transcript-api/>). Automatic Speech Recognition (ASR) transcripts were chosen to filter out non-English language videos.

2. Literature Review

2.1 Corpus cleaning

According to Kehoe (2020), web-based text can often be 'noisy', featuring a lack of punctuation or whitespace, and irregular spellings. In addition, he points out that part-of-speech taggers can be less accurate when tagging web-based texts, as they are usually

trained on well edited, written texts.

2.2 Part-of-speech tagging accuracy

State of the art part-of-speech taggers generally have around 97% per token accuracy, which is similar to human annotator agreement (Manning, 2011). According to Jurafsky and Martin (2022) most part-of-speech taggers are evaluated by accuracy, and specific annotations such as named-entities are evaluated by recall, precision, and F-measure. To calculate precision, it is necessary to know the occurrences of a feature tagged as a feature (true positives) and words other than the feature that have been tagged as the feature (false positives). For recall, true positives and the occurrences of a feature not tagged as the feature (false negatives) are needed, and the F-measure combines precision and recall (Picoral, Staples & Reppen, 2021). If any pre-processing is likely to affect the accuracy of the part-of speech tagging, its accuracy should be evaluated. Two research questions were posed:

1. What features should be cleaned in YouTube ASR transcripts?
2. How accurate are the proposed pre-processing solutions?

3. Methodology

3.1 Identifying features for cleaning in the corpus

To identify features that should be cleaned, the YouTube plain text file transcripts were examined manually, and regular expressions were used in Python to check lines that had punctuation, to check if there were any upper-case words, and to check the words that included square brackets. Examining punctuation is useful to check for sentence boundaries, and other text features such as hyphenated words and apostrophes. It was found that there were no full stops, except occasionally following whole numbers, such as ‘i’m 19. my name is...’, ‘no that’s 2018. it’s not 2018...’, meaning there were no sentence boundaries. The transcripts were basically all lower case, except for occasional occurrences of ‘T’, and some other exceptions such as one occurrence of ‘Messi’. Upon checking, the MAT Tagger tagged lower case ‘i’ as a foreign word, which also appears to have caused tagging problems in other corpora. For example, in the Open American National Corpus (spoken) which is available on Sketch Engine (<https://www.sketchengine.eu/>), 93% of the occurrences of the word ‘i’ were tagged as a proper noun. Non-capitalised proper nouns were also not tagged correctly, therefore, it seems that capitalisation substantially affects the accuracy of tagging. Finally, YouTube ASR transcripts feature words in square brackets that do not represent speech, specifically ‘[Music]’, ‘[Applause]’, and ‘[Laughter]’. In addition, the following characters

‘[_]’ represent words that are censored in the transcript, but not in the audio.

3.2 The cleaning process

The following features were deleted using regular expressions: any periods that followed numbers (excluding decimal points), and words that did not represent speech (e.g. ‘[Music]’). The following features were amended: ‘i’ was capitalised, and censored words were amended to the pseudo word *tamn*. This was chosen based on the word ‘damn’. Although damn is not censored in YouTube videos, it has the function of a swear word. Using the EnTenTen20 corpus in Sketch Engine, it was found that damn was tagged as an adverb, adjective, interjection, noun and verb. Two other frequent swear words, ‘shit’ and ‘fuck’ were only tagged as a noun or verb.

To capitalise proper nouns, the English caseless model from the Stanford Tagger version 4.2.0 was used (<https://stanfordnlp.github.io/CoreNLP/caseless.html>). As the MAT Tagger accepts only plain text files, all transcripts were tagged, then the proper nouns were capitalised, then tags were removed from the transcripts. This process is not perfect, for example, ‘U.S.’ (United States) appeared in the original transcript as ‘us’, so following the process, the word was restored to ‘Us’. This process took around one hour to process 4.4 million words of text. The final problem was punctuation. Several tags are only applied following punctuation by the MAT Tagger, such as sentence relatives (e.g. they are sugar free, *which* is nice). In addition, having no punctuation in the text could cause problems for the part of speech tagger syntactically. To overcome this problem, a punctuation prediction model (Guhr et al, 2021) was used to restore full stops, commas and question marks. Hyphens and colons predicted by the model were replaced with commas. The process took about 10 hours to process around 4.4 million words of text.

3.3 Calculating accuracy

The accuracy of the Stanford Tagger English caseless model was evaluated by randomly sampling 20 transcripts (38,394 tokens) from a corpus of YouTube transcripts containing 2,682 texts. All occurrences of proper nouns and instances of a proper noun tag being applied were searched, and true positives, false positives and false negatives were recorded. Then the accuracy, recall and F-scores were calculated for each text.

The accuracy of replacing censored words with the word *tamn* was evaluated by checking all of the tags applied to *tamn* by the Stanford Tagger version used in the MAT Tagger in a random sample of 20 transcripts that featured censored words (196 observations of *tamn*). The ‘gold standard’ tag was the tag that should have been applied to the original word in the video. The metric of accuracy was used by calculating the percentage of words that had the correct tag applied to them.

It is difficult to evaluate the accuracy of punctuation in texts that did not feature

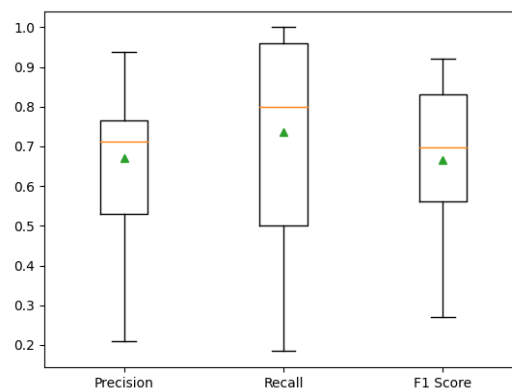
punctuation originally, as punctuation is used somewhat subjectively. However, examples of how one text was punctuated are reported below.

4. Results

4.1 Proper noun tagging accuracy

The mean F-score for proper noun tagging accuracy for texts that were tagged with the Stanford caseless model was 0.67 ($SD = 0.18$). The distribution of mean text scores for precision, recall and F-scores is shown in Figure 1.

Figure 1 *Proper noun tagging accuracy by the Stanford caseless model*



4.2 Censored word tagging accuracy

Correct tags were applied to the censored word replaced by *tamn* 66.3% of the time. For individual texts, the mean accuracy was 55.7% ($SD = 34.2$), but this data was skewed by texts that only featured one or two censored words.

4.3 Punctuation prediction

Here are some examples of how the punctuation restoration was applied to a portion of one video (<https://www.youtube.com/watch?v=iRxeX5PL9I>). To begin with, only full stops were applied. However, this meant that many sentences started with fillers or interjections:

um just ask how how frustrating that was. uh you know that you dominated long periods but you had to wait until the end to get something out of the game.

Here are some examples of occasions where hyphens and colons were restored:

we we didn't show any signs of a hangover

it's as simple as that: to give us ourselves the best chance of qualifying.

It was felt that as hyphens and colons are used inconsistently, and to a certain extent subjectively that they could be replaced by commas. The following is an extract with all punctuation restored and hyphens and colons replaced by commas:

um, just ask how, how frustrating that was. uh, you know that you dominated long periods, but you had to wait until the end to get something out of the game. yeah, really frustrating. um, i think we started the game well.

Question marks also seem to be restored accurately: ‘um, does that give you cause for optimism ahead of tuesday?’

5. Discussion

The accuracy of the Stanford caseless model at tagging proper nouns might not seem particularly high (F-Score = 0.67, $SD = 0.18$). However, when we compare it with the MAT Tagger F-Score (0.01, $SD = 0.03$), it seems that this process is worth doing. From the box plots in Figure 1, we can see that the mean accuracy scores for individual transcripts are quite highly dispersed. This is probably due to the wide range of textual domains in YouTube videos. For example, the texts with the highest accuracy scores had topics such as Christianity and basketball. These texts may have been more similar to the domains of texts the tagger was trained on compared with texts with lower scores. One example of a video with a lower score was a video about Pokémon cards, which contained many names of Pokémon characters that were not recognised. On other occasions, acronyms were tagged as proper nouns and other mistakes were caused by ASR errors in the transcripts.

Similarly, the accuracy of replacing all censored words with the pseudo word *tamn* may not seem very high (66.3%). The most frequent error was ‘fucking’ tagged as a noun (e.g. ‘my fucking_NN eyebrows_NNS’, ‘the fucking_NN drink_NN’) when it should have been tagged as an adjective. Other errors were caused by the morphological properties of the word, such as plural nouns tagged as singular nouns. An alternative solution would be to manually check all of the censored words and amend the tags manually. However, this would be time consuming, as the only way to find the original word is to listen to it in the video. Censored words occurred fairly frequently. In a corpus of 2,682 videos (4.4m words), there were 2,222 censored words in 237 videos.

The suitability of the punctuation prediction model has only been evaluated by checking a small number of transcripts and judging whether the model should be used. The following F-scores are reported by the model creators (Guhr et al, 2021) for full stops (0.94), question marks (0.89), commas (0.81), colons (0.58), hyphens (0.43), and no punctuation (0.99). The model was trained on the Europarl parallel corpus, which contains multilingual transcripts from the European Parliament Proceedings, therefore accuracy could be different in domains such as YouTube videos. However, as the short examples in this paper show, the output seems to be suitable for spoken texts, where it

is not always clear where punctuation should be added.

6. Conclusion

As part of a research project to choose appropriate YouTube videos for EFL learners, the steps taken to pre-process YouTube transcripts were summarised. To further improve the tagging of proper nouns, a caseless model could be trained on YouTube transcripts. However, the performance of the caseless model 'out-of-the-box' already greatly improves accuracy when compared with the Stanford Tagger version used by the MAT Tagger. The problem of censored words is a difficult one to solve, and it seems like there is no simple solution, unless the researcher can check the texts, and audio manually. The difference in the performance of the MAT Tagger before and after punctuation restoration could also be evaluated further by comparing texts with and without that form of pre-processing. A number of concrete methods have been described that will hopefully be useful for other researchers interested in doing research with spoken language and specifically YouTube transcripts. The Python code used in this paper is available at https://github.com/cooperchris17/YouTube_transcript_preprocessing.

Bibliography

- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- Guhr, O., Schumann, A.-K., Bahrmann, F., & Böhme, H.-J. (2021). FullStop: Multilingual Deep Models for Punctuation Prediction. *Proceedings of the Swiss Text Analytics Conference 2021*. Retrieved from http://ceur-ws.org/Vol-2957/sepp_paper4.pdf
- Jurafsky, D. and Martin, J. (2022). *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd Ed. Draft). Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf
- Kehoe, A. (2020). Web Corpora. In M. Pacquot & S. T. Gries (Eds.) *A Practical Handbook of Corpus Linguistics* (pp. 329-351). Springer. https://doi.org/10.1007/978-3-030-46216-1_15
- Manning, C.D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In A. F. Gelbukh (Ed.) *Computational Linguistics and Intelligent Text Processing* (pp. 171-189). Springer. https://doi.org/10.1007/978-3-642-19400-9_14
- Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues* (67-94). Bloomsbury Academic. <http://dx.doi.org/10.5040/9781350023857.0012>
- Picoral, A., Staples, S., & Reppen, R. (2021). Automated annotation of learner English: An evaluation of software tools. *International Journal of Learner Corpus Research*, 7(1), 17-52. <https://doi.org/10.1075/ijlcr.20003.pic>

日英・英日パラレルコーパス検索ツール『パラレルリンク』(Ver.1.20) ーインターフェース, 検索機能, 活用研究などについてー

仁科 恭徳(神戸学院大学 教授／英国オックスフォード大学 客員研究員)

ynishina@gc.kobegakuin.ac.jp

赤瀬川 史朗(Lago NLP 代表)

shiro.akasegawa@lagonlp.jp

"Parallel Link" (Ver. 1.20), a tool for searching Japanese-English
and English-Japanese parallel corpora: About the interface,
search functions, and research applications

NISHINA Yasunori (Kobe Gakuin University, Professor/ Oxford
University, Visiting Research Scholar)
AKASEGAWA Shiro (Lago NLP, CEO)

Abstract

This presentation details the interface, search functions, and application research of the online Japanese-English and English-Japanese parallel corpus search tool "Parallel Link" (Ver. 1.20). Nishina and Akasegawa (2021, 2022) detailed the specific procedures applied to nine existing parallel corpora. This presentation particularly describes the follow-up procedures: dependency analysis, content word patterning (nouns, verbs, adjectives, adverbs, and collocations), lexical profile database creation and so on.

Keywords

パラレルコーパス, 日英・英日翻訳, コーパス言語学, レキシカルプロファイラー, ツール開発

1. はじめに

本研究では、日英・英日パラレルコーパスオンライン検索ツール『パラレルリンク』(Ver.1.20) のインターフェース、検索機能、活用研究などについて詳細に紹介する。仁科・赤瀬川 (2021、2022) では既存の 9 種のパラレルコーパスに施した具体的な処理を詳説したが、本発表ではその後に施した係り受け解析、内容語の文法パタン作成 (名詞、動詞、形容詞 (イ形容詞、ナ形容詞)、副詞、連体詞)、レキシカルプロファイリングデータベースの作成 (内容語ごとの共起関係のデータ (共起語、頻度、MI スコア、LogDice 等)) について説明する。また、レキシカルプロファイリングツール開発の背景についても紹介する。

2. 『パラレルリンク』(Ver.1.20) のインターフェースと検索機能

プロトタイプとなる『パラレルリンク』(Ver.1.20)については、試行的に新インターフェースを搭載したが、色々と改善点も見つかったため、無償一般公開を目指す上でも、NINJAL-LWP for BCCWJ (NLB) や NINJAL-LWP for TWC (NLT)のインターフェースを改良したものを、この Ver.1.20 で搭載することにした。

まず、画面構成については、検索タブに見出し語パネル(品詞ごと)を、見出し語タブに文法パターンパネル、コロケーションパネル、用例パネルをそれぞれ配置している。これらのパネルは、レキシカルプロファイリングのデータベースの見出し語情報テーブル、文法パターンテーブル、コロケーションテーブル、用例テーブルからそれぞれデータを取得している。見出し語情報テーブルとは、見出し語(漢字かな交じり、ひらがな、ローマ字)、品詞、頻度、活用形別頻度(用言のみ)など、見出し語に関する基本情報を収録したテーブルを指す。また、文法パターンテーブルとは、見出し語ごとに、文法パターン別の頻度、パタングループにおける割合などを収録したテーブルを指す。そして、コロケーションテーブルとは、見出し語と文法パターンをキーとして、該当する共起語と統計値(単純頻度、MI スコア、LogDice 等)を収録したテーブルを指す。最後に、用例テーブルとは、見出し語と文法パターンとコロケーションをキーとして、該当する用例とその出典(ページタイトル、URL)を収録したテーブルを指す。このあたりの説明も実際にインターフェースを提示しながら説明する。

Home 画面の右上には「検索を開始する」ボタンが配置されている。このボタンをクリックすると、検索タブに移動する。Ver.1.20 では、日→英方向のみ検索が可能であることから、現状、日本語の各見出し語の情報のみが一覧表示される。予め用意されたタブは「すべて」、「名詞」、「動詞」、「イ形容詞」、「ナ形容詞」、「連体詞」、「副詞」である。検索したい品詞のタブをクリックすれば、頻度順で各品詞に該当する語の情報が手に入る。検索ボックスへの文字入力、かな漢字交じりのほか、ひらがな、カタカナ、ローマ字による検索も可能にしている。正規表現を使った検索コマンドも用意しているので、紹介したい。

参考までに、図 1 は見出し語検索画面を、図 2 は検索ボックスで「立てる」を指定した画面を、図 3 はカタカナで「タテル」を指定した画面を示す。図 4 は正規表現を用いて、3文字のタテルに該当する異種のみを拾い上げている検索画面である。図 5 は「立てる」の語彙プロファイルを示す。図 5 の語彙プロファイルから、左側にはパターンパネルがあり、中央にはコロケーションパネルが配置されており、右側にはジャンル(サブパラレルコーパス)ごとに頻度情報と対訳が表示されていることが分かる。

見出し検索

調べたい語を入力

絞り込み 元に戻す

すべて 名詞 動詞 イ形容詞 ナ形容詞 連体詞 副詞

見出し	読み	頻度
する	スル	367,973
いる	イル	358,444
ある	アル	325,243
こと	コト	261,369
なる	ナル	203,313
その	ソノ	189,402
この	コノ	186,449
年	ネン	185,292
条	ジョウ	168,181
いう	イウ	152,395
よる	ヨル	148,131
項	コウ	139,700
もの	モノ	133,656
ない	ナイ	119,861
場合	バアイ	114,181
できる	デキル	108,666
よう	ヨウ	90,381

1 of 729 1 - 100 of 72803

図 1 見出し語の検索画面

見出し検索

立てる

絞り込み 元に戻す

すべて 名詞 動詞 イ形容詞 ナ形容詞 連体詞 副詞

見出し	読み	頻度
立てる	タテル	4,190
申し立てる	モウシタテル	292
組み立てる	クミタテル	251
積み立てる	ツミタテル	245
取り立てる	トリタテル	221
仕立てる	シタテル	164
見立てる	ミタテル	138
役立てる	ヤクダテル	117
駆け立てる	カリタテル	86
打ち立てる	ウチタテル	74
引き立てる	ヒキタテル	57
追い立てる	オイタテル	37
埋め立てる	ウメタテル	34
突き立てる	ツキタテル	29
騒ぎ立てる	サウギタテル	25
責め立てる	セメタテル	17
言い立てる	イイタテル	16
用立てる	ヨウダテル	14
泡立てる	アワダテル	13
逆立てる	サカダテル	12

図 2 検索ボックスで「立てる」を指定した場合

見出し検索

タテル

絞り込み 元に戻す

すべて 名詞 動詞 イ形容詞 ナ形容詞 連体詞 副詞

見出し	読み	頻度
立てる	タテル	4,190
たてる	タテル	442
申し立てる	モウシタテル	292
組み立てる	クミタテル	251
積み立てる	ツミタテル	245
取り立てる	トリタテル	221
仕立てる	シタテル	164
見立てる	ミタテル	138
駆り立てる	カリタテル	86
かきたてる	カキタテル	78

図 3 検索ボックスで「タテル」を指定した場合

見出し検索

↑タテル\$

絞り込み 元に戻す

すべて 名詞 動詞 イ形容詞 ナ形容詞 連体詞 副詞

見出し	読み	頻度
立てる	タテル	4,190
たてる	タテル	442
点てる	タテル	33
裁てる	タテル	1

図 4 検索ボックスで「↑タテル\$」を指定した場合

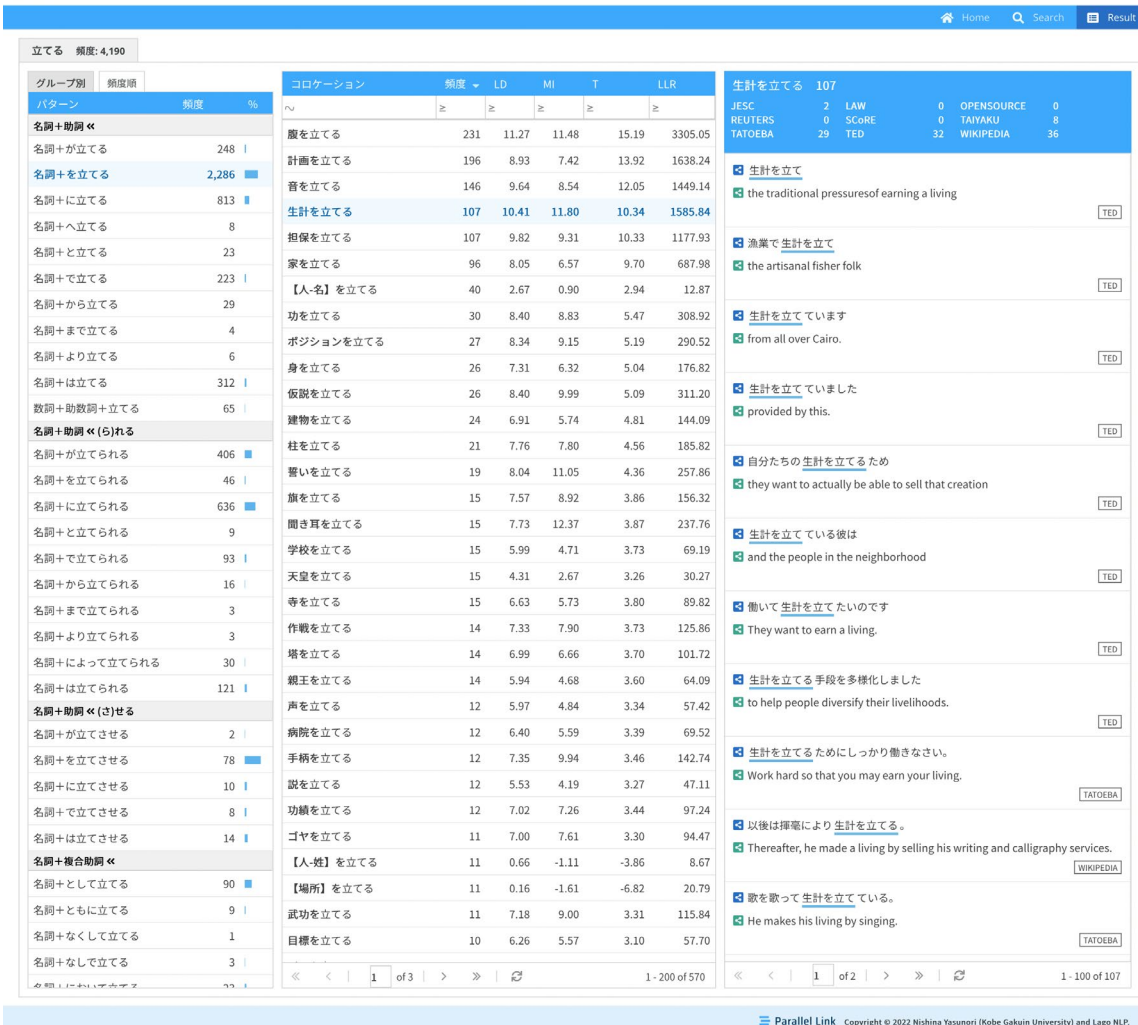


図 5 「立てる」の語彙プロフィール

3. まとめ

以上のように、本発表では、『パラレルリンク』(Ver.1.20)のインターフェース、検索機能などを実際のインターフェースを画面に表示しながら紹介する。また、仁科(2023)でも触れている想定される活用研究についても実例を挙げて紹介したい。

謝辞

本研究は JSPS 科研費 20K00692 の助成を受けたものである。また、神戸学院大学海外研究員制度を活用することで、英国で研究に没頭することができた。ここに感謝申し上げる。そして、Score の用例コーパスを搭載することをご快諾くださった中條清美先生(元日本大学教授)、現在 Score の一連の研究を引き継いでおられる西垣知佳子先生(千葉大学教授)、パラレルコーパスの重要性を再認識してくれた故・染谷泰正先生(元関西大学教授)、ならびに全ての関係者の皆様に感謝の意を示す。

引用文献

- 仁科恭徳(2023)『パラレルコーパス言語学の諸相:モノリンガルコーパス研究からバイリンガルコーパス研究へ』 開拓社.
- 仁科恭徳・赤瀬川史朗(2021)「日英・英日パラレルコーパスオンライン検索ツール『(仮称)パラレルリンク(Ver.1.0)』の開発に向けて(中間報告)」『英語コーパス学会大会予稿集 2021』, 25-30.
- 仁科恭徳・赤瀬川史朗(2022)『パラレルリンク』(Ver.1.0)の開発:パラレルコーパス研究の概観とコーパス整備』『英語コーパス研究』, 29, 63-78.

受動文における様態副詞の生起位置に関する一考察

西村 知修(石川工業高等専門学校)

s13ac041tomomichi@gmail.com

On Manner Adverb Placement in the Passive

NISHIMURA Tomomichi

(National Institute of Technology, Ishikawa College)

Abstract

The basic position of adverbs in the passive is considered to be in front of the past participle although they can also be positioned postverbally. According to previous research, modified noun phrases, such as *a carefully examined donor*, affects adverb placement in the passive, which leads to a preference for the adverb-initial order, such as *be carefully examined*. Through an examination of examples from COCA, this study reveals that some manner adverbs are frequently placed behind the past participle even in the passive. It also shows that the influence of modified noun phrases on adverb placement is not immediately clear in some cases.

Keywords

様態副詞, 受動文, 動作受動, 状態受動

1. はじめに

様態副詞は基本的に、(1)の能動文、(2)の受動文のどちらにおいても動詞の前後に生じることが可能である。能動文では(1a)のように動詞の後ろ側が基本的な位置とされ、受動文では(2b)のゆに過去分詞の前側が基本的な位置であるとされる。以下、例文の強調は筆者による。

- (1) a. In January of this year, the North Carolina state attorney general ***examined carefully*** all the evidence. (COCA 2007: SPOK)
b. Drew ***carefully examined*** the body of the young boy. (COCA 2013: FIC)
- (2) a. But it needs to ***be examined carefully*** as a last-ditch alternative to the dangers of living with an Iranian bomb. (COCA 2009: ACAD)
b. Kai ***was carefully examined*** and measured. (COCA 2002: NEW)

また動詞の必須要素となる様態副詞は、受動文では動詞の前側に生起が制限されるという指摘を

安藤(2005)がしている。

(3) *She **was treated badly**.

(安藤 2005: 525)

これらの指摘を合わせると、能動文の動詞を様態副詞が修飾する場合は動詞の後ろ側に生起すること、受動文の場合は動詞の前側に生起することが基本であるというのは揺るぎのない事実のように思える。様態副詞に限らず副詞が受動文で中間位置に高い頻度で現れるのは事実のようだが、動詞と副詞の組み合わせによってその頻度は異なり、(3)で非文とされる **be treated badly** も実例では散見される。本研究では受動文における副詞の位置に影響する要因として、〈副詞＋過去分詞形＋名詞〉の語順が影響を与えているという先行研究を概観し、それが様態副詞にも当てはまるか検討し、いくつかの様態副詞と動詞の組み合わせではその影響は確認できないと指摘する。

2. 先行研究

Pijpops and Van de Velde(2016)は、語順にバリエーションがある構文 A に対して、その構文の語順バリエーションの 1 つと表面上同じ語順を持つ別の構文 B が高い頻度で現れる場合、構文 A の語順が構文 B の語順に影響を受けることがあることをオランダ語の構文で例証している。これを英語の受動文の副詞の位置に援用したのが Hilpert and Flack (to appear) である。彼らは〈副詞＋過去分詞形＋名詞〉の語順が受動文における副詞が生じる位置に影響を与える可能性を検証している。(2)の例を用いて具体的に述べるなら、受動文では **be carefully examined** と **be examined carefully** の語順がありうるが、例えば **a carefully examined donor** のような例に多く接することにより、受動文でも **be carefully examined** の語順になる傾向が強まるというのである。Hilpert and Flach はそのような影響があるということを統計的に示している。

本研究は統計的な手法は用いないが、〈副詞＋過去分詞形＋名詞〉の語順が受動文における副詞の位置に影響する可能性を探る点は共通している。しかし受動文の過去分詞形が形容詞とみなされて、〈副詞＋過去分詞形＋名詞〉の影響と関係なく副詞が動詞の前に生じる可能性がある。安藤(2005: 525)は受動文での副詞の語順が〈副詞＋形容詞〉と同じ語順となることから、「受動分詞が英語母語話者にとって形容詞と同様に受け取られていることの一つの証拠となる」と述べている。受動文におけるどのような過去分詞を実際に形容詞とみなすかについては様々な主張があるが、少なくとも受動文が動作受動と状態受動で曖昧になる場合があることは事実である。

(4) The window **was broken**.

(Huddleston and Pullum 2002: 1439)

(4)は「壊された」という行為を表す場合と「壊れていた」という状態を表す場合がある。例えば様態副詞は基本的には状態を表す述語を修飾しない。

(5) *He **likes** them **skillfully**.

(岡田 1985: 44)

能動文では動詞の後ろ側に生じるのが基本の様態副詞は、状態受動の文には生起することさえできず、〈副詞＋過去分詞形＋名詞〉の影響を受けるまでもなく、受動文における副詞の生起位置は動詞の前側に偏る。このように考えると、本当の意味での〈副詞＋過去分詞形＋名詞〉の影響を検討するためには、動作受動のときの副詞の位置を観察する必要がある。曖昧性なく動作受動が現れるのは(6)のように動作主を示す *by* 句がある場合や、(7)のように動詞が *kiss* や *kick*, *hit* などの状態変化を表さない働きかけ動詞の場合(影山 2007: 73), また(8)のような *get* 受身などの場合であるとされる。

(6) The kitchen window ***was broken*** by the thieves.

(Huddleston and Pullum 2002: 1436)

(7) The man ***was hit*** twice.

(COCA 2012: BLOG)

(8) The window ***got broken***.

(Huddleston and Pullum 2002: 1441)

また〈副詞＋過去分詞形＋名詞〉の影響を分析するためには、能動文の場合に動詞の後ろ側に生起するのが基本である様態副詞のような副詞の位置をより詳細に観察する必要があるように思われる。例えば *also* や *often*, *only* などの副詞は能動文でも動詞の前側によく生じるため、受動文でも普通に動詞の前側に生じることが予測され、〈副詞＋過去分詞形＋名詞〉の影響がわかりにくい。

3. リサーチデザイン

前節において、受動文の〈副詞＋過去分詞形〉と〈副詞＋過去分詞形＋名詞〉の関係をより詳細に分析するためには、動作受動の文において副詞がどの位置に生起するのか、また能動文では動詞の後ろ側に生起するのが基本である副詞が受動文でどこに生起するのかを観察する必要があることを述べた。そこで動作受動の解釈になる可能性が大きいと考えられる *kiss*, *kick*, *hit* と共起する副詞、また様態副詞の解釈しかないとされる中核的純様態副詞(Schäfer 2002)の *loudly*, *tightly*, *brightly* と共起する動詞をそれぞれ調査する。また受動文では動詞の前側に生起が限定されるという指摘がある *treat* と *badly* の組み合わせも調査をする。この調査のために COCA (オンライン版)を用いた。まずはターゲットとなる動詞 3 語(*kiss*, *kick*, *hit*)と副詞 3 語(*loudly*, *tightly*, *brightly*)の左右 1 語範囲の共起語を検索し、それらと共起する高頻度の副詞・動詞を選んだ。ただし副詞は *-ly* 副詞から選び、さらに頻度を表す副詞なども除外した。分析の対象とする語の組み合わせを決めたのち、例えば *kick* と *wildly* の組み合わせであれば, [be] *wildly kicked*, [be] *kicked wildly* で検索して受動文の副詞の位置とその数を、また *wildly kicked* で検索して〈副詞＋過去分詞形＋名詞〉の数を、最終的に目視で確認した。

4. 結果と考察

前節で述べた調査を行い、*kiss*, *kick* については以下のような結果になった。*hit* についてはあ

る程度の生起数がある副詞の共起語として適当な-ly 副詞がなかったため割愛する。

表 1 kiss, kick と共起する副詞の生起位置

副詞	過去分詞	副詞＋過去分詞	過去分詞＋副詞	副詞＋過去分詞＋名詞
passionately	kissed	0	1	0
gently	kissed	0	0	0
lightly	kissed	2	0	0
wildly	kicked	0	0	0
violently	kicked	1	2	0
furiously	kicked	0	0	0

能動文の kiss や kick には passionately, gently, lightly はある程度の数が共起するが、受動文で共起する例は表 1 の通りほとんどなかった。また lightly kissed の 2 例を見ると、(9)のように kiss は「キスをする」の意味ではなく、また lightly も「わずかに」という程度の意味だと考えられる。

- (9) The rich crab is supported by a layer of lightly seasoned, ripe avocado that *is lightly kissed* with Meyer lemon juice to prevent the avocado from darkening.

(COCA 2004: NEWS)

また lightly や violently は能動文でも高い頻度で動詞の前側に生じるため、受動文でも動詞の前側に生じることは不思議ではない。これらのわずかな例のみからは断定できないが、副詞の受動文での生起位置は能動文での生起位置と深い関係があり、〈副詞＋過去分詞形＋名詞〉が影響を与えていると言えるだけの証拠はない。

次に loudly, tightly, brightly と共起する動詞と、それらの動詞と副詞との位置関係の結果を見てみよう。それぞれの副詞と said の共起例が多く見られたが、語順が特殊なので除外した。また自動詞の用法が主で受動文の例が非常に少ない過去分詞形(sighed)と loudly の例も除外した。

表 2 loudly, tightly, brightly と共起する動詞と副詞の生起位置

副詞	過去分詞	副詞＋過去分詞	過去分詞＋副詞	副詞＋過去分詞＋名詞
loudly	laughed	0	0	0
loudly	complained	0	0	0
loudly	called	2	0	0
tightly	controlled	111	16	206
tightly	held	6	22	23
tightly	wrapped	14	31	28
brightly	colored	29	0	1227

brightly	lit	66	2	411
brightly	painted	10	1	185

まず loudly に関して, laughed と complained と共起する例は検索できなかった。called と共起する場合は called の前側に loudly が生じる例が見られた。〈loudly+called+名詞〉は見つからなかったので, この例においては〈副詞+過去分詞形+名詞〉が受動文の副詞の位置に影響を与えたとは言えない。

次に tightly に関して, まずは controlled との共起を考察する。そもそも tightly は能動文のときでさえ動詞 control の前側に生じるのが普通であり, 受動文でも動詞の前側に生じることは不思議ではない。tightly が hold や wrap と共起する場合は純粋に物理的様態を表しているが, control と共起する場合はより比喩的な意味となり, それが一因となって能動文でも動詞の前側に生じると考えられる(西村 2020: 188)。よって tightly が controlled と共起する場合, 受動文の場合に動詞の前側により多く生じるようになったと考えるのは誤りで, 能動文と受動文のどちらでも動詞の前側が基本の位置なのである。〈tightly+controlled+名詞〉の頻度が高いので, この語順の存在が受動文における〈be+tightly+controlled〉の語順を強化している可能性はあるが, 受動文〈be+tightly+controlled〉の語順を決定する一番の要因と言えるかは不明確である。一方で tightly が hold や wrap と共起する場合は物理的様態を表し, さらに held や wrapped が動作受動を表すことがあるため, held tightly や wrapped tightly の語順が優勢なのだと考えられる。tightly と動詞の共起に関してまとめると, tightly が純粋な様態副詞で動作受動の文に生じる場合, 動詞の後ろ側に生じる傾向がある。また tightly が純粋な様態用法でなく, そもそも能動文においても動詞の前側により生じる傾向がある場合, 〈副詞+過去分詞形+名詞〉の語順の影響を受けるまでもなく, 受動文においても動詞の前側に生じると考えられる。

次に brightly に関して, 共起する過去分詞はほとんどが状態受動である。状態受動は形容詞に近いので, brightly が状態受動過去分詞の前側に生じることは妥当である。brightly のように純粋な様態副詞であっても, 状態受動の文に生起するときは動詞の前側に置かれるということである。〈brightly+過去分詞+名詞〉の生起数が非常に多いため, この語順が受動文の brightly の位置に影響を与えている可能性は否定できないが, 受動文における brightly の位置を決める主要因だとは言えないと思われる。

最後に badly と treated の共起に関する結果である表 3 を見てみよう。

表 3 treated と共起する badly の生起位置

副詞	過去分詞	副詞+過去分詞	過去分詞+副詞	副詞+過去分詞+名詞
badly	treated	45	91	0

先行研究の指摘に反して be treated badly の生起数のほうが多いことが判明した。実際以下のよう、文を成立させるうえでの必須要素の badly が treated の後ろに生起している例が多数ある。

(10) A circus employee at a recent performance denied the animals *were treated badly*.

(COCA 2010: NEWS)

様態副詞はそれが文の必須要素であろうとなかろうと、動作受動の文では過去分詞の後ろ側に生じる傾向があると考えられる。また〈badly + treated + 名詞〉の例を検索することはできなかったにもかかわらず、受動文で〈be + badly + treated〉の語順もある程度の生起数がある。ここでも〈副詞 + 過去分詞形 + 名詞〉は受動態の副詞の位置を決める決定的な要素とは言えない。

5. まとめ

受動文における副詞の生起位置に関して、先行研究では過去分詞の前側が基本の生起位置であるとされている。また〈副詞 + 形容詞 + 名詞〉の語順が受動文における副詞の位置に影響を与える可能性が議論されている。しかし COCA で受動文におけるいくつかの様態副詞の生起位置を調査した結果、必ずしも過去分詞の前側に生じる例ばかりではないことが判明した。特に純粋に物理的様態を表す様態副詞が動作受動の文に生じた場合は、むしろ過去分詞の後ろ側により多くの副詞が生起することが示された。また〈副詞 + 形容詞 + 名詞〉の影響が必ずしも明らかではない場合も示された。今後より多くの副詞と動詞の組み合わせを調査する必要がある。

引用文献

安藤貞雄 (2005) 『現代英文法講義』開拓社.

Hilpert, M. and S. Flach. (to appear). A case of construction contamination in English: Modified noun phrases influence adverb placement in the passive. In M. Grygiel (ed.), *Contrast and analogy in language: Perspectives from cognitive linguistics*. John Benjamins.

Huddleston, R. and G. K. Pullum (2002). *The Cambridge grammar of the English language*. Cambridge University Press.

影山太郎 (2007) 「get 受身文の統語構造と概念構造について」『英米文学』(関西学院大学), 51(2), 63-82.

西村知修 (2020) 「様態副詞の基本位置とそこからの逸脱」八木克正・神崎高明・梅咲敦子・友繁義典(編)『英語実証研究の最前線』179-193, 開拓社.

岡田伸夫 (1985) 『副詞と挿入文』大修館書店.

Pijpops, D. and F. Van de Velde. (2016). Constructional contamination: How does it work and how do we measure it? *Folia Linguistica*, 50(2), 543-582.

Schäfer, M. (2002). Pure manner adverbs revisited. *Sinn und Bedeutung VI, Proceedings of the Sixth Annual Meeting of the Gesellschaft für Semantik*, 311-323.

起動表現の意味と補部にくる語との関係性に関する考察 —start NP を例に—

藏菌 和也(神戸学院大学)

kurazono@human.kobegakuin.ac.jp

On the Relation between the Syntax and Semantics of Inchoative Expressions —With Special Reference to Start NP—

KURAZONO Kazuya (Kobe Gakuin University)

Abstract

In this paper, we investigate what kind of nouns frequently occur in the complement of *start* and consider why certain kind of nouns are chosen as complement noun.

First of all, the previous research on the choice of noun in the complement of *start* are previewed and a theory about selectional restrictions on its complement is constructed. Secondly, the validity of the theory is tested by conducting a quantitative and qualitative survey on the complement nouns of *start* by means of the British National Corpus and a comparative analysis of *start NP* and *begin NP* is provided.

The result shows that *start* frequently takes the nouns such as engine, company, team, and business expressing the entities which autonomously move or develop after someone or something made them activated. It is safe to say that the causality and instantaneity of *start* have a great effect on the choice of nouns occurring in its complement.

Keywords

起動動詞, 動詞の補文, 名詞の選択, 語彙の意味と統語

1. はじめに

起動動詞と呼ばれる出来事の開始を表す動詞の中でも *start* が補部に名詞をとる場合(以降, *start NP* と呼ぶ)には, 特有のコロケーションが見られる。先行研究でも, 同じ起動動詞 *begin* と比較してみると, 補部において(1)のような名詞選択の違いが見られることが説明されている。また, 小西(編)(1980)は(2)のようなコロケーションを示して *start* が「(事業など)を起こす[設立する]」「妊娠する」といった意味を表すことを説明している。本発表では, *start* の補部にどのような名詞がくるかに関して動詞の意味の観点から考察を行う。

(1) a. Bob started/ *begin the car. (小西(編)1980: 1492-6) (Freed 1979: 80)

- b. When are you going to start/*begin the fire? It's getting cold in here.
- c. If you keep at it, you are going to start/*begin a fight.
- d. Do you have any idea who started/*began the rumor? (ibid.: 80)
- (2) a. He decided to start newspaper. (新聞を始める)
- b. He (=Freud) started psychoanalysis. (精神分析学の創始者)
- c. start a baby 「妊娠する」 (小西(編)(1980: 1493))

2. 先行研究

Freed(1979)では, **start** は(i) (2)のように使役性(causative)を持つ動詞であることから, (2c)や(3)のように使役構文での言い換えが可能で, (3)のような構造をとる時には補文, ここでは **thinking** の主語が文主語と異なってもよいこと, (ii) **start** が(5a)のように車やエンジン(motor or moving part)の「運動」の意味を表すこと, さらに(5b)のように使役者の影響で出来事が開始する段階(causative onset of an event)が表されること, (iii) (6)のように動名詞よりも派生名詞が好まれるという性質が見られること, (iv)コロケーションに関する先行研究には, 冒頭(1)や(2)で示した研究のほかに, The Corpus of Contemporary American English (以降, COCA と呼ぶ)を利用した量的な調査を行った中村(2018)が見られる。この研究では, **start+a/an** に後続する名詞には(7)のようなものが, **start+the** に後続する名詞には(8)のようなものが, それぞれ後続しやすいと言う。また, 不定冠詞をとる場合と冠詞をとる場合でも, **start NP** に頻出する名詞は「(点)火, 始動」と「組織, 事業, グループ」に関するものである(中村 2018: 55)と言う。

- (2) a. He started the painting yesterday.
- b. The painting was started (by him) yesterday.
- c. He got the painting started yesterday
- . (Cf. *He got the painting begun yesterday.) (Freed 1979: 78-79)
- (3) a. Joe got me started thinking about the problem.
- b. Joe caused me to start thinking about the problem. (ibid.: 80)
- (4) Joe started me thinking about the problem. (Cf. *Joe began me thinking about the problem.) (ibid.: 79)
- (5) a. Bob started/ *begin the car. (= (1a)を再掲) (Cf. Bob **made** the car started.)
- b. The explosion started/ *began a landslide. (Cf. The explosion **made** a landslide start.) (ibid.: 80)
- (6) The district attorney began/started the questioning of the defendant. (ibid.: 83)
- (Cf. The district attorney began/started questioning the defendant.)
- (7) a. **fire** が最も **begin** との差異を表す表現で, 着火比喩的に車を始動させることも表す名詞 (fire, car, car engine)
- b. 組織やグループを意味する名詞群(team, band)

- c. 事業や計画としてまとめられる表現 (company, business, agency, shop, factory など)
- (8) a. 車や機械に関わる表現 (engine(s), car, motor, truck, machine, ignition, jeep, van)
- b. ignition「(エンジンに) 点火すること」と共通する「点火する」の意味の表現 (fire(火・発砲・火災の意の場合も含む), blaze)
- c. 組織や事業を意味する名詞 (company, foundation)

3. リサーチデザイン

3.1 研究の目的

これまでの研究を振り返ると, **start** は補部に様々な名詞句をとり, 多様な出来事の開始を表すことが分かる。ある時には, 車やその動力部であるエンジンが「運動」つまり動き始めることを表し, 時に火 (**fire**) が燃え始める様子を表す。また, ある時には喧嘩や戦い (**fight**) が始まったり, 噂 (**rumor**) が出回り始める様子が表現される。このように多様な名詞句が **start** の補部にくる事実が先行研究の記述から確認できたが, 補部にはどのような性質の名詞句が生起するのか, 言い換えると, どのような意味特性を持つ名詞句が生起するかについての説明は十分ではないように思える。

例えば, 小西(編) (1979) では **start the fire/a fight/the rumor** という用例をそれぞれ挙げて特有のコロケーションで特有の意味を表すことを示しているが, それぞれの語がなぜ **start NP** に生起するのか, 各語がどのような性質が共通するのかを説明するといったことは行われていない。

本発表では, 起動動詞 **start** の補部に生起する名詞句に何らかの共通性は見られるのか, さらにその背景には動詞 **start** のどのような意味が関係しているのかを明らかにすることを目的とし, コーパスを用いて **start** の補部に生起する語の量的および質的調査を行った。そして, 調査の結果から得られたデータを基に議論を進めていく。

3.2 データ

まず, **start NP** に生起する名詞句の中でも典型的に使用される名詞コロケーションを抽出するために, 大規模汎用コーパスを利用した。コーパスは, 米英語モニターコーパス COCA を利用した調査(中村 2018)が既に存在すること, さらに英米の英語において **start** の補部で典型的に利用される名詞句について違いがあった場合を考慮して, イギリス英語の話し言葉および書き言葉を約 1 億語を収集した大規模汎用コーパス The British National Corpus(以降, BNC と呼ぶ)を利用してデータを収集した。検索には, Sketch Engine のコーパス検索言語(CQL)機能を用いて行った。

3.3 手法

3.3.1 データの事前処理

検索には, 検索式 ([lemma="start"] [tag="NN.|AJ.|AV.|DPS|PN.|AT0|NN1|NN2"] {0,3}[tag="N.*"])を利用して **start** の右 3 語以内に生起する名詞を抽出した。3 語以内としたのは, (10)のように冠詞や形容詞など補部の名詞の前に何らかの語がくる場合を考慮して用例を収

集するためである。検索の際に、(11)のような start NP とは関係ない例も含まれていたため、それらの例を手作業で除外し、生起頻度上位 20 語までの単語を表 1 にまとめた。

(10) During the same period Mr Benson got up at nine, read his mail, started his office work at ten, and worked through till lunch at one. (BNC)

(11) a. Housing scheme starts </s><s> WORK has started on a major business and housing development in north Essex. (BNC)

b. Her account in effect exonerates her father, who was now starting a life sentence for shooting her husband. (BNC)

3.3.2 分析の手順

分析手順は、先行研究でネイティブスピーカーが start の補部に用いた場合に容認性が異なると説明する名詞を観察し、どのような共通性が見られるのか理論を提示した後に、コーパスのデータを基に start の補部に生起する名詞を観察し、その理論が妥当かどうかを検討した。

表 1 BNC にみる動詞 start の補部に生起する名詞

順位	名詞	頻度	順位	名詞	頻度
1	work	480	11	campaign	60
2	life	242	12	job	56
3	school	127	13	game	53
4	family	116	14	season	42
5	day	111	15	programme	41
6	career	110	16	war	40
7	engine	109	17	proceedings	39
8	business/fire	102	18	conversation/operation	32
9	course	76	19	project	31
10	process	75	20	year	30

4. 結果と考察

4.1 補部にくる名詞とその特徴からの分類

表 1 を見てみると、動詞派生の名詞 (conversation, operation) は少数派である。名詞を大まかに分類してみると、(i) 仕事や学校、生活に関する語 (work, life, school, career, business, course, process, job, production, programme, operation, project), (ii) 意見交換や対立に関する語 (conversation, proceedings, fight, war, game (試合に出場)), (iii) 機械やその一部を表す語 (engine), (iv) 期間に関する語 (day, season, year), (v) その他 (family (第一子が産まれる), fire, campaign) のように分類できそうだ。しかし、この分類から名詞の共通性を理解するの

は難しく、start が補部に要求する名詞の性質について考えてみる必要があるように思える。

4.2 典型的な名詞コロケーションとその性質

Freed (1979) で類義語である begin の補部では容認性が低い start the car/landslide/fire/fight/rumor というコロケーションを詳細に観察してみると、自動車(car)はエンジン(engine)が始動するとエンジンを止めるまで自力で動くという性質がある。次に、地滑り(landslide)は、土砂が崖などの斜面を平地に至るまで重力の力を受けて落ちていくという性質がある。さらに火(fire)は点火するとその後は木やガスが燃え尽きるまで自力で燃え続けるものである。喧嘩(fight)についても対決が始まるとその勢いを増していきやがて決着が着き終わりを迎えるものである。最後に、噂(rumor)というのも人から人へと伝わり、事実が分かるまで広まっていく性質を持つ名詞である。これまで述べた通り、Freed が挙げている start の補部にくる特有の名詞の性質を考えると、開始のために力が加わることで、その後、何らかの力を借りながら(車やエンジンなどのように)自力で動いたり、火などのように力を増したり、喧嘩のように展開したり、噂のように広まる性質、つまり名詞が表すものが自ら動きを与えられて動いたり展開していく性質がみられる。Freed が start は causative onset of the event を表すと言うように、start は名詞で表されるモノが何らかの影響で(=使役性)、静の状態から動いている状態へと変化する(=開始の意味)様子を描いる。このように start が使役の意味と開始点の意味を表す語彙的意味を持っているので、start NP における car/landslide/fire/fight/rumor の使用の容認性が高いものと考えられる。

4.3 高頻度で使用される名詞の考察と理論の検証

前述のように「自律的な出来事の展開が予想できる」という性質は、start の補部に生起するほかの名詞にも共通する。先行研究で start との結びつきが強いとされる company, shop, team, band などはいずれも人が経営したりリーダーとして指揮を執り、その経営およびチーム内の状況は時間と共に自ずと変化して組織は成長していくといった展開が予想できる名詞である。期間を表す名詞も、一日(day)が始まれば慌ただしく過ぎ去っていく。また、仕事や事業(work, career, business, campaign, job, project など)が始まったり学校(school)に入学すると仕事や授業、学校の課程が終わるまで決められたコース(course, programme)を進んでいくものである。さらに、試合(game)に出場した後もゲームは展開していくという性質の出来事が名詞から予想できる。ただ、名詞が表す出来事の開始から展開までを意識して表しているのではなく、開始の時点を表すことが、一人目の子供が生まれることや妊娠し始めたことを表す start a family/baby という表現からも推測できる。

4.4 begin NP との比較から見ると特有の意味とコロケーション

藏菌(2021)では begin NP の補部に生起する名詞の中でも生起頻度上位 10 語にくる語を調査し、work, life, career, process, series, campaign, operation, investigation, study, discussion などの名詞がくることが分かった。さらに、11 位から 20 位までの語を本研究で調査すると proceedings, tour, talk, journey, visit, task, search, programme, descent, day,

sentence, preparation などの名詞が見られた。この結果と start NP に生起する名詞上位 20 位を比較すると、例えば(12)の例に代表されるように begin NP には一定期間継続する出来事を表す名詞、つまり出来事の過程を意識させる名詞が使用される(藏菌 2021)。このことが、モノが動き始める際の開始点に着目する start NP とは異なる特有のコロケーションが見られる大きな要因となっていると考えられる。(12)は全て藏菌(2021)が挙げている BNC からの用例である。

- (12) a. However, brief mention should be made here of the Indian press. The first of these papers was Indian Opinion, which began its long career in 1906, launched by Mahatma Gandhi.
- b. A team from the human rights organization Amnesty International arrived on Dec. 4 to begin a two-week investigation into political killings.

5. まとめ

本発表の目的は、起動動詞 start の補部にどのような性質の名詞句が生起するかについて量的・質的調査を行い、start the car/newspaper/a baby のような特有のコロケーションが形成される背景に見られる意味的動機を解明することであった。

BNC を用いて起動表現 start NP に高頻度で生起する名詞句を量的に調査し、さらに質的調査として補部に生起する名詞句の意味を分類し、共通性について考察を行った。その結果、start の補部では「起動後に自律的に動き、発展するモノ」を表す名詞が選択されやすいことが分かった。また、このような共通性が補部に生起する名詞句に見られる背景に、start が持つ行為の開始点に焦点を当てる性質 (instantaneity) と使役性 (causality) が深く関係していることを説明した。

謝辞

※本研究は、JSPS 科研費 JP21K13029 の助成を受けたものである。

引用文献

- Freed, Alice (1979) Semantics of English Aspectual Complementation, D. Reidel, Dordrecht.
- 藏菌和也.(2021).「起動動詞の補部に関する研究—begin NP を例として—」『日本英語コミュニケーション学会紀要』第 30 巻第 1 号, 37-51.
- 小西友七.(編).(1980).『英語基本動詞辞典』東京:研究社.
- 中村文紀.(2018).「コーパスに基づく類義語分析—start/begin+名詞を例に一」『北里大学一般教育紀要』第 23 巻, 49-61.

コーパス

BNC: The British National Corpus (2020 年 10 月に検索)

日本人中学生英語学習者の *take* の使用と語義指導への提言

澤口 遼(関西大学第一中学校)
rswgch@gmail.com

The Use of *Take* by Japanese Junior High School Learners and Practical Suggestions for Vocabulary Teaching

SAWAGUCHI Ryo (Kansai University Dai-ichi Junior High School)

Abstract

Different senses of the verb *take* are introduced early in the Japanese junior high school English learning curriculum. However, little attention has been given to the teaching and presentation order of the senses according to learners' proficiency. Hence, this study explores how Japanese junior high school learners in the JEFLL corpus acquire different meanings of the senses of the verb *take* as their learning progresses. Analysis using EVP showed that (1) learners acquire a wider range of senses as they advance in their grades, and (2) they have semantic preferences for senses of the verb *take*. Based on these findings, I propose practical suggestions for teaching.

Keywords

CEFR, EVP, 語彙知識の深さ, 語彙指導

1. はじめに

英語における基本語かつ多義語でもある *take* の語義の多くは中学校で導入されるが, 学習者の習熟度に応じてどの段階で各語義を提示・指導すべきかは十分に検討されていない。語彙指導はカリキュラム開発者・指導者の主観的判断だけでなく, 学習者の習熟度やニーズを勘案し行われるべきである(Tono & Aoki, 2000)。そこで本研究では, 英語学習者コーパス Japanese EFL Learner Corpus(以下, JEFLL)(投野, 2007)の中学生学習者が使用する *take* の語義を外国語能力指標 CEFR(Council of Europe, 2001)レベル別語義データベース English Vocabulary Profile(以下, EVP)(Capel, 2015)によって分類し, 中学1年から3年までの各学年でどの CEFR レベルの語義が使用可能になるのか調査することを目的とする。続いて, 彼らの発表語彙知識の発達を考慮した語義の指導法について考察する。

2. 先行研究

語彙指導は外国語教育において重要な項目の 1 つであり、学習者の語彙習得の実態を解明し、得られた知見を指導や評価に活用する動きが高まっている(Hawkins & Filipovic, 2012)。特に、多義語に関して学習者の発表語彙の運用に関して有益な情報を提供しているのが EVP(Capel, 2015)である。EVP は世界各国の異なる CEFR レベルの学習者が習熟度に応じて語のどの語義が使用可能になるのかを記述したデータベースであり、例えば同じ **make** であっても、**make lunch**(昼食を作る)という原義に近い平易な意味であれば A1 レベル(英検 3 級程度)の学習者が使用可能な語義で、**make a difference**(変化を起こす)という派生的意味であれば B2 レベル(英検準 1 級程度)の学習者が使用可能な語義であることがわかる。このような EVP の特性は、学習者の発表語彙知識の発達の観察に適していると言える。

Barker(2015)や Leńko-Szymańska (2015)は各国の英語学習者の使用する語と EVP を比較し、学習者の習熟度を予測、評価することを試みている。しかしながら、これらの研究では日本において **take** の語義の学習を本格的に開始する日本人中学生学習者が **take** の語義をどれだけ知っており、その語彙知識が学習の進行と共にどのように発達していくかという発表語彙の質的深化については記述が少ない。**Take** は英語において高頻度かつ多様な語義を持つ語であり、日本人学習者の語義習得を EVP という客観的枠組みで記述することは教育的意義があると考えられる。

3. リサーチデザイン

3.1 研究目的と研究設問

本研究の目的は、英語学習者コーパス JEFLC の中学生学習者が使用する **take** の語義を EVP によって分類し、中学 1 年から 3 年までの各学年でどの CEFR レベルの語義が使用可能になるのか調査することである。前章の先行研究の課題をふまえ、本研究では以下 2 点の研究設問を設定した。

RQ1: 日本人中学生英語学習者の **take** の語義使用率は学年とともにどのように変化するか

RQ2: 日本人中学生英語学習者が使用する **take** の語義にはどのような意味的傾向があるか

3.2 データ

本研究では日本人中学生英語学習者の **take** の語義使用における発表語彙知識の発達を調査するため、英語学習者コーパス JEFLC を使用する。JEFLC は日本人中高生 1 万人のデータを集めた大規模学習者コーパスであり、日本人中高生学習者の発表語彙知識発達の傾向を観察するのに有用であると考えられる。また、トピックも朝食の嗜好や童話の創作など幅広く、学習者が英語で表現できなかった箇所は日本語の使用を認めているため、より自発的な学習者の発達語彙知識の使用を捉えることが可能という特徴を持つ。サブコーパスの総語数は中学 1 年が 51,149, 中学 2 年が 159,736, 中学 3 年が 117,766 語である。

3.3 手法

JEFLLL の「サブコーパス」機能によって中学生学習者(J1・J2・J3)にデータ収集を限定し、レマ化した **take** を起点とするコンコーダンス検索によって、抽出された **take** の語義使用を調査する。収集した語義の CEFR レベルへの分類には、EVP を用いる。なお、EVP にはイギリス英語(British)とアメリカ英語(American)の語義が提示されているが、日本の学校英語教育ではアメリカ英語が主流であるため、本研究でもアメリカ英語の語義を参照し、議論を進める。2 章で述べた通り、CEFR は A→B→C とアルファベット順に習熟度が上昇するが、同一のレベルでも 1 よりも 2の方が習熟度は高いとされる。したがって、A1 よりも A2 の方が、B1 よりも B2 の方が習熟度が高い。図 1 は EVP の A1 レベルの **take** の語義検索画面である。JEFLLL の中学生学習者が(1)のような表現を使用していた場合、その学習者は図 1 の語義 **GO WITH SOMEONE** が使用可能であると判断する。コロケーションなどの誤りや EVP に記載がなかったことにより語義が認定できなかった **take** の使用は分析から除外し、中学1年で 15, 中学 2 年で 198, 中学 3 年で 269 の CEFR レベル別語義を認定した。また、学習者の使用する語義のほぼ全てが A1・B1 の範囲に位置していたため、これら 3 つの語義を分析対象とした。

(1) I want to take them to [JP:hinan_basho] If I can do it. (j2_earthquake_01778)

図 1 EVP の語義検索画面

take · verb [T]  /teɪk/

+ Word family

+ take (CARRY)

A1 to get and carry something with you when you go somewhere

- take (GO WITH SOMEONE)

A1 to go somewhere with someone, often paying for them or being responsible for them

Dictionary examples:

*I took the kids **to** the park.*

*I thought I'd take her **to** that nice Italian restaurant.*

Learner example:

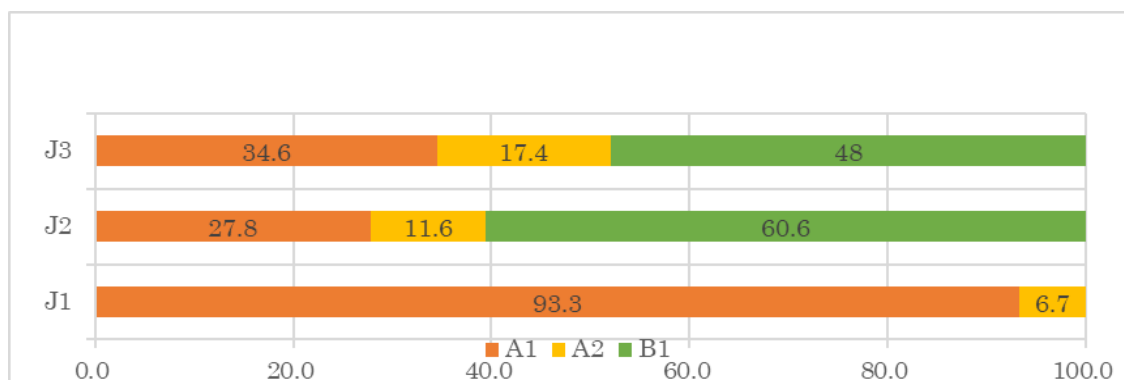
She takes her son to school. (Skills for Life (Entry 1): A1; Turkish)

4. 結果と考察

4.1 RQ1 学年ごとの take の語義使用率の変化

中学 1 年から 3 年までの学習者の take の CEFR レベル別語義が全ての語義に占める割合を求めたところ、図 2 の結果を得た。

図 2 学年別語義使用率(%)



結果をもとに、以下の 2 点を考察する。1 点目は、語義の使用が A1 レベルにほぼ限定され、B1 レベルの語義使用が皆無である中学 1 年の学習者と比較し、中学 2 年・中学 3 年の学習者は A1 レベルの語義の使用が減少し、A2 レベルの語義の使用の増加とともに、語義使用率のおよそ半数を B1 レベルの語義が占めるようになるという点である。これは学年の上昇に伴い、学習者がより高度かつ多様な語義を使用するようになり、語彙知識が質的に深化していくことを示唆している。2 点目は、学習者の使用する語義には偏りがあるという点である。中学 2 年・中学 3 年の学習者は B1 レベルの語義を全ての語義中で最も多く使用している一方、A2 レベルの語義の使用率は 3 つの語義中で最も低いという傾向がみられた。このことは学習者の使用語義の選好性を表している。

4.2 RQ2 学習者の語義使用の意味的傾向

中学 1 年から 3 年までの学習者が使用する語義のうち、頻度の高いものを CEFR レベル別に上位 3 種類まで計上した結果が表 1 である。なお、サブコーパスサイズが異なる学習者間の比較のため、各語義の使用頻度は 10 万語あたりに調整した。表 1 の各語義の名称は EVP の分類に基づく(一部略)。

表 1 学年別 10 万語あたりの高頻度使用語義

J1	J2	J3
picture	12 take out	68 take out 103
CARRY	10 CARRY	19 TIME 31
GO	4 TIME	10 CARRY 30

興味深い点は、中学 1 年の学習者が最も頻繁に使用する語義は A1 レベルの **take a picture** (写真を撮る)であるのに対し、中学 2 年・中学 3 年の学習者は B1 レベルの **take out** (物を持ち出す)という語義を最も多く使用しているという点である。JEFL には「大地震が起きたら何を持って避難するか」というトピックが含まれるため、それが中学 2 年・中学 3 年の学習者の **take out** の高頻度使用につながったと考えられるが、彼らの中学校での言語使用到達指標は概ね A2 程度である(投野, 2013, p.160)ことを考慮すると、比較的高度な **take** の語義使用が可能であると言えるだろう。また、中学 3 年の学習者が 2 番目・3 番目に多く使用する語義はほぼ同一の頻度で A2 レベルの **TIME**(何かを行うのに時間を要する), A1 レベルの **CARRY**(物を持っていく)であり、彼らが中学 1 年・中学 2 年の学習者と比較して多様な語義をバランスよく使用していることが窺える。

母語や該当表現との親密度の観点では、母語と同一で「写真を撮る」と直訳可能な **take a picture** や、日本語に浸透している「テイクアウト」が学習者の習得を促進している可能性がある。こうした傾向は Negishi, Tono & Fujita(2012)でも言及がある。

また、**take out** については、結合可能語に一定の範囲を持つコロケーション(Nesselhauf, 2003, p.226)(**take/do/sit an exam** など)よりも、**take** と **out** において語彙的結束性の強い **take out** は、学習者にとって固定化されたチャンクとしてメンタルレキシコンへの格納が容易である可能性を示唆している。

上記の点をふまえると、中学校における **take** の語義指導では学習者の習熟度に応じた語義の導入や指導を検討することが必要となるだろう。例えば、中学校の検定済英語教科書 **SUNSHINE ENGLISH COURSE 2** では中学 2 年で **take a look** という B1 レベルの軽動詞的語義が導入されているが、本研究の結果から、中学 2 年の多くの学習者はその語義が発表語彙として習得できる可能性が高いと推測できる。この点をふまえ、指導生徒の習熟度を鑑みた上で、小テストで該当語義を出題する、語義を使用する機会を確保するためのアクティビティを行うなど、発表語彙として定着させるための活動を考案することは効果的であろう。教材開発の観点では、中学 2 年と中学 3 年においては積極的に B1 レベルの語義を導入する一方、その 1 つ前の段階である A2 レベルの語義の記述も充実させ、中学校で導入される **take** の語義の段階的な定着を図ることが重要であると考えられる。

5. まとめ

本研究では、日本人中学生英語学習者の **take** の語義の使用が学習段階とともにどのように変化していくのか、語義の使用にはどのような意味的傾向があるのかという 2 点について、EVP を用いて調査した。その結果、学年が上がるにつれ、学習者はより多様な CEFR レベルの語義が使用可能になるが、使用する語義には意味的偏りがあるということが明らかとなった。英作文トピックの影響が存在するため、結果の解釈には慎重な姿勢が必要である。しかし、本研究は学習者の語彙知識の深さを EVP の枠組みで捉えるという点で語彙習得研究に新たな手法を提供し、実践面では学習者の習熟度に応じた指導や教材開発の手法を検討するための一助となることが期待される。

引用文献

- Barker, F. (2015). Portraits of each CEFR level. In J, Harrison. & F, Barker (Eds.) *English profile studies*, 5, 69-92. Cambridge University Press.
- Capel, A. (2015). The English vocabulary profile. In J, Harrison. & F, Barker (Eds.) *English profile studies*, 5, 9-27. Cambridge University Press.
- Council of Europe. (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Hawkins, J. A. & Filipovic, L. (2012). *Criterion features in L2 English*. Cambridge University Press.
- Leńko-Szymańska, A. (2015). The English Vocabulary Profile as a benchmark for assigning levels to learner corpus data. In Callies, M., & Götz, S (Eds.) *Learner corpora in language testing and assessment*, 115-140. John Benjamins Publishing.
- Negishi, M., Tono, Y. & Fujita, Y. (2012). A validation study of the CEFR level of phrasal verbs in the English vocabulary profile. *English Profile Journal*, 3(1), 1-16.
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223-242.
- Tono, Y., & Aoki, M. (2000). A computer-based analysis of the acquisition order of irregular verbs. *KATE Bulletin*, 14, 103-122.
- 投野由紀夫(編著)(2007)『日本人中高生一万人のコーパス: JEFLL Corpus』小学館.
- 投野由紀夫(編著) (2013)『英語到達度指標 CEFR-J ガイドブック』大修館書店.

文部科学省検定済中学校英語教科書

『SUNSHINE ENGLISH COURSE 2』(2020)開隆堂.

学習者コーパスにおける品詞連鎖の分析による母語の影響に関する研究—名詞句構造に焦点を当てて—

山口 一華(東京外国語大学 大学院生)
yamaguchi.ichika.w0@tufs.ac.jp

投野 由紀夫(東京外国語大学)
y.tono@tufs.ac.jp

L1 Influence on the Acquisition of English Noun Phrases —Analyzing Part-of-Speech Sequences—

YAMAGUCHI Ichika (Tokyo University of Foreign Studies,
Graduate Student)
TONO Yukio (Tokyo University of Foreign Studies)

Abstract

This study analyzes the part-of-speech sequences of English noun phrases in learners' corpora with different L1s (Japanese and German) with the aim of obtaining insights into the L1 influence on second language acquisition. This study especially focuses on noun phrases used in subject position. By attaching information of syntactic dependencies along with part-of-speech tags makes it possible to take the syntactic structure into account. The results shows that Japanese learners tend to underuse articles and overuse compound nouns, and the learners whose L1 is German start to use prepositional phrases at a very early stage of the acquisition.

Keywords

Second language acquisition, crosslinguistic influence, learners' corpora, Part-of-Speech sequences, noun phrases

1. はじめに

第二言語習得における母語の影響に関する研究は、これまで第二言語習得理論に関わる多くの研究者の関心の中心となり、多岐に渡る実証研究が行われてきた(Jarvis and Pavlenko, 2008)。近年、学習者コーパスを含む新しい方法論の確立により、今までは行うことのできなかった大規模かつ多面的な定量分析が可能になり、学習者の産出言語(中間言語)の体系的な記述研究が行われるようになった(Granger, 2017)。中間言語の大まかな特徴を把握するための研究の一つとして、学習者コーパス内で使用されている品詞(POS)タグの連鎖を調べる研究が挙げられる(Aarts

& Granger, 1998、Tono, 2000)。本研究では、そのような研究で得られた成果を考慮し、母語の影響に関してさらなる示唆を得るため、統語構造を加味した品詞連鎖を分析する。特に、本研究は日本語・ドイツ語をそれぞれ母語に持つ英語初級～中級学習者が主語の位置で用いる名詞句の品詞連鎖に着目した分析を行う。

2. 先行研究

2.1 品詞連鎖を用いた研究

学習者の中間言語のおおまかな傾向をつかむ目的で品詞 (POS) タグの連鎖を調べた先駆的な研究として、Granger (1998) が挙げられる。Granger (1998) では、統計文体学の分野で、作者推定のために用いられていたタグ連鎖を定量する手法から着想を得て、3 種の学習者コーパスと母語話者コーパスにおける品詞タグの 3 個組 (trigram) の抽出を行った。結果として、各学習者グループの中間言語には、学習者に普遍的な言語特徴より、L1 特有の言語特徴の方が大きい割合を占めることが明らかになった (約 7% vs 20%)。Granger (2017) では、この研究手法の有用性を強調すると共に、より具体的な言語特徴に着目した研究や統語構造を加味したさらなる研究の必要性が述べられている。さらに Tono (2000) では、異なる発達段階にいる学習者の中間言語の記述を目的に、日本語を母語とする初級～中級学習者の産出言語を同様の手法を用いて分析した。結果として、冠詞や助動詞の過少使用などを観察した。また、Granger (1998) と Tono (2000) に共通して見られた結果として、学習者の中間言語における前置詞の過少使用が挙げられる。特に名詞を後置修飾する前置詞句の使用の困難性は、L1 に関わらず、学習者に共通してみられる特徴であることがこれらの研究によって示された。

3. 研究方法

3.1 研究目的と研究設問

本研究の目的は、母語の異なる英語学習者コーパスを用いて、学習者が使用する品詞連鎖を分析し、第二言語習得における母語の影響に関して示唆を得ることである。本研究では特に、ドイツ語母語話者と日本語母語話者の名詞句使用に着目した分析を行う。また、今回は学習者が産出した言語から名詞句のみを正確に抽出する目的で、文の主語として用いられる名詞句を対象に分析を行う。名詞句使用の難しさはこれまでの研究ですでに明らかにされているが、ここでは、ドイツ語の名詞句構造が、英語と比較的類似していることを指摘しておきたい。以下の(1)では、形容詞が名詞を前置修飾した名詞句、(2)では、前置詞句が名詞を後置修飾した名詞句、(3)では形容詞、前置詞句などの修飾句が名詞に付加する複雑な名詞句の例を挙げている。

- (1) diese schöne Blume (this beautiful flower)
- (2) eine Katze unter dem Tisch (a cat under the table)
- (3) viele schwierige Bücher über das Land (many difficult books about the country)

から(3)で紹介した名詞句すべてが、その英語訳と完全に語順が一致している。このことから、ドイ

ツ語を母語とする英語学習者にとって、複雑な名詞句の使用は容易である可能性が考えられる。これらを踏まえ、「名詞句における品詞の使用とその発達過程に母語の影響はみられるか」という研究設問を設定した。

3.2 データ

本研究では、日本語を母語に持つ英語初級～中級学習者として JEFLL コーパス (投野, 2007) を、ドイツ語を母語に持つ英語初級～中級学習者として ICCI コーパス (Tono, 2012) のオーストリアサブコーパスを使用する。JEFLL コーパスとは、延べ 1 万人の日本人中高生が産出した自由英作文データを 10 年間以上の長期間にわたって収録したものであり、データの総語数は、約 60 万語に及ぶ。ICCI コーパスは、JEFLL コーパスとの比較を目的に、対象をその他のアジア各国・中東・ヨーロッパの初級～中級学習者に広げた国際的な学習者コーパスである (投野 et al., 2013)。本研究では、ICCI コーパスのうちドイツ人学習者のデータは収録されていないため、ドイツ語を母語とするオーストリア国籍の学習者サブコーパスを使用する。

3.3 分析手順

3.3.1 データの事前処理

本研究では、データの事前処理として、主語にあたる名詞句の抽出が困難になるような文 (疑問文や母語を含んだ文、ピリオドが抜けている文など) をあらかじめ削除した。その上で、Python の自然言語処理のライブラリである stanza (Qi, Zhang et al., 2020) を利用した構文解析を行い、学習者の言語使用データを、統語依存関係と品詞情報を示すタグの連鎖に変換した。タグは「:」で区切り、「統語依存関係 (DEPREL):品詞情報 (UPOS)」を記載した。使用したタグおよび品詞一覧は表 1 のとおりである。

表 1 使用したタグセット

nsubj:pron 主語名詞 (代名詞)	nmod:poss:pron 名詞修飾の代名詞
nsubj:noun 主語名詞 (名詞)	det:det 冠詞
nsubj:propn 主語名詞 (固有名詞)	amod:adj 限定形容詞
expl:pron 嘘辞 (仮主語の it や there)	compound:noun 複合等位接続詞の名詞
flat:propn 同格複数単語表現:固有名詞	nmod:noun 名詞修飾の名詞
case:adp 動詞修飾の前置詞	
cc:cconj 等位接続詞	

3.3.2 分析の手順

事前処理を行ったデータから、本動詞として使用されている動詞を基準に、主語にあたる名詞句の抽出を行った。その際に主語である名詞句に前置する前置詞句や副詞句は、統語依

存関係をもとに削除した。また、分析を複雑にするのを避ける目的で、受動態を示すタグや副詞を示すタグなどを削除した。その上で、学習者の母語と学年ごとに、統語依存関係と品詞情報を示すタグの連鎖の頻度を数え、上位 10 位までをランキング形式で出力した。

4. 結果と考察

本研究の分析対象になった文の総数と、名詞句の総語数は表2と表3の通りである。また、母語と学年ごとの頻度に基づいた上位 10 位までのタグの連鎖を以下の表 4(日本語母語話者)と表 5(ドイツ語母語話者)で示した。

表2 JEFLL コーパスの基礎統計情報

	J1	J2	J3	S1	S2	S3
センテンス数	3949	13027	9982	7165	16099	3060
単語数	5142	17508	13173	9969	24651	4782

表3 ICCI コーパス JEFLL コーパスの基礎統計情報

	G5	G6	G7	G8	G9	G10	G11
センテンス数	365	742	1243	1151	1000	498	413
単語数	525	1078	1869	1655	1648	751	826

表 4 JEFLL コーパスにおける主語を表す名詞句のタグ連鎖(上位 10 位)

J1	J2	J3
nsubj:pron	3037 nsubj:pron	9657 nsubj:pron
nmod:poss:pron nsubj:noun	206 nmod:poss:pron nsubj:noun	620 nmod:poss:pron nsubj:noun
nmod:poss:pron compound:noun nsubj:noun	124 det:det nsubj:noun	298 expl:pron
det:det nsubj:noun	93 nsubj:propn	227 det:det nsubj:noun
nsubj:propn flat:propn	71 nmod:poss:pron compound:noun nsubj:noun	226 nsubj:propn
nsubj:propn	64 nsubj:noun	187 nsubj:noun
nsubj:noun	46 nsubj:propn flat:propn	157 nsubj:propn flat:propn
amod:adj nsubj:noun	26 expl:pron	148 nmod:poss:pron compound:noun nsubj:noun
compound:noun nsubj:noun	16 amod:adj nsubj:noun	111 amod:adj nsubj:noun
expl:pron	15 nsubj:propn punct:punct flat:propn	48 subj:propn punct:punct flat:propn
S1	S2	S3
nsubj:pron	5214 nsubj:pron	10970 nsubj:pron
nmod:poss:pron nsubj:noun	303 nmod:poss:pron nsubj:noun	601 det:det nsubj:noun
nsubj:noun	198 det:det nsubj:noun	501 nmod:poss:pron nsubj:noun
det:det nsubj:noun	166 expl:pron	491 expl:pron
expl:pron	158 nsubj:noun	368 nsubj:noun
amod:adj nsubj:noun	85 nmod:poss:pron compound:noun nsubj:noun	188 nsubj:propn
nsubj:propn	81 nsubj:propn	185 amod:adj nsubj:noun
nmod:poss:pron compound:noun nsubj:noun	68 amod:adj nsubj:noun	170 nmod:poss:pron compound:noun nsubj:noun
nsubj:propn flat:propn	53 det:det amod:adj nsubj:noun	77 nsubj:propn flat:propn
det:det compound:noun nsubj:noun	21 vocative:propn punct:punct nsubj:pron	68 det:det amod:adj nsubj:noun

表 5 ICCI コーパスにおける主語を表す名詞句のタグ連鎖(上位 10 位)

G5	G6	G7
nsubj:pron 251	nsubj:pron 544	nsubj:pron 901
nmod:poss:pron nsubj:noun 51	nmod:poss:pron nsubj:noun 79	nmod:poss:pron nsubj:noun 60
det:det nsubj:noun 17	nmod:poss:pron amod:adj nsubj:noun 23	det:det nsubj:noun 59
nmod:poss:pron nsubj:noun cc:cconj conj:pron 8	det:det nsubj:noun 17	nsubj:pron 21
nmod:poss:pron amod:adj nsubj:noun 7	expl:pron 6	expl:pron 20
nmod:poss:pron nsubj:noun cc:cconj nmod:poss:pron conj:noun 3	nsubj:noun 5	nmod:poss:pron amod:adj nsubj:noun 17
root:adv 3	obj:pron nsubj:pron 4	det:det amod:adj nsubj:noun 12
det:det compound:pron nsubj:pron 2	nmod:poss:pron nsubj:noun cc:cconj conj:pron 3	nmod:poss:pron nsubj:noun cc:cconj conj:pron 8
nmod:poss:pron nsubj:noun cc:cconj conj:noun 2	nsubj:pron cc:cconj nmod:poss:pron conj:noun 3	nsubj:pron det:det 6
det:det nsubj:noun case:adp nmod:poss:pron nmod:noun 1	nmod:poss:pron compound:noun nsubj:noun 2	nsubj:pron flat:pron 6
G8	G9	G10
nsubj:pron 890	nsubj:pron 683	nsubj:pron 371
nmod:poss:pron nsubj:noun 47	det:det nsubj:noun 50	nmod:poss:pron nsubj:noun 18
det:det nsubj:noun 31	nmod:poss:pron nsubj:noun 40	expl:pron 17
expl:pron 23	expl:pron 25	det:det nsubj:noun 13
nmod:poss:pron amod:adj nsubj:noun 10	nsubj:pron 19	nmod:poss:pron amod:adj nsubj:noun 6
nsubj:noun 8	nmod:poss:pron amod:adj nsubj:noun 13	nsubj:noun 6
nsubj:pron nsubj:pron act:relcl:verb 8	nsubj:noun 9	det:det amod:adj nsubj:noun 3
det:det amod:adj nsubj:noun 7	det:det amod:adj nsubj:noun 8	det:det nsubj:noun advmod:pron 3
nmod:poss:pron nsubj:noun cc:cconj conj:pron 5	nmod:poss:pron nsubj:noun cc:cconj conj:pron 7	det:det nsubj:noun case:adp nmod:pron 2
nsubj:det 5	nummod:num nsubj:noun 4	mark:part 2
G11		
nsubj:pron 179		
nsubj:noun 37		
expl:pron 33		
amod:adj nsubj:noun 24		
det:det nsubj:noun 22		
det:det amod:adj nsubj:noun 6		
nmod:poss:pron nsubj:noun 6		
det:det nsubj:noun case:adp nmod:noun 4		
advmod:adp fixed:noun nsubj:pron 2		
compound:noun nsubj:noun 2		

4.1 RQ 母語の影響

ここからは、研究設問「名詞句における品詞の使用とその発達過程に母語の影響はみられるか。」についての考察を行う。まず、日本語母語話者の中間言語に着目すると、ドイツ語母語話者に比べ冠詞を使用した名詞句が少ないことが特徴として挙げられる。これは、これまでの研究で明らかにされている日本人学習者の冠詞の過小使用の特徴と重なる結果である。また、日本人学習者は複合名詞を使用した名詞句を好んで用いているようである。この傾向の原因として、日本語の名詞句構造の影響が考えられる。一方で、ドイツ語母語話者は日本語母語話者に比べ、英語習得初期段階から等位接続詞などを積極的に使用し、複雑な統語構造を使用する傾向にある。日本語母語話者の頻度上位 10 位では見られなかった前置詞句の使用も、日本の中学一年生に相当する学年である第 5 学年から見受けられた。このことから、名詞の後置修飾句は母語にかかわらず一様に難しいわけではなく、類似した言語構造が存在する言語を母語に持つ学習者にとっては、比較的容易に習得される可能性を示唆する。

5. まとめ

本研究は、第二言語習得における母語の影響に関する示唆を得ることを目的に、母語の異なる英語学習者の言語データにおいて主語にあたる言語表現の品詞連鎖を分析した。結果としては、

母語の影響が考えられる中間言語の特徴として、日本人学習者に冠詞の過少使用や複合名詞の過剰使用がみられ、ドイツ語母語話者には前置詞句の使用が見受けられた。

しかし、本研究では、主語にあたる名詞句のみに対象を絞った分析を行ったため、主語にみられる特徴を反映した結果になった可能性が大いに考えられる。今後は目的語にあたる名詞句など、他の統語構造にも目を向けた研究を行いたい。また、本研究では、学習者の産出言語のみに着目し、目標となる英語母語話者または上級英語学習者のベンチマークとなる名詞句連鎖を分析に取り入れることができなかった。母語話者の産出言語について考慮しない分析は、母語の影響に関する研究を進めるにあたって、不十分であるといえる。最後に、*stanza* の構文解析精度が学習者の初級レベルでどの程度正確であるかは検証が必要である。今後は、このような改善点について考慮してさらなる分析を進めたい。

引用文献

- Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger, *Learner English on Computer* (1st ed., pp. 132–141). Addison Wesley Longman.
- Granger, S. (2017). *Learner English on Computer* (1st edition). Routledge.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. <https://doi.org/10.18653/v1/2020.acl-demos.14>
- Tono, Y. (2000). A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora. Lewandowska-Tomaszczyk, B. and J.P. Melia, (eds.) *PALC'99: Practical Applications in Language Corpora*. Peter Lang. pp.323- 340
- Tono, Y. (2012). International Corpus of Crosslinguistic Interlanguage: Project Overview and a Case Study on the Acquisition of New Verb Co-occurrence Patterns. In Tono, Y., et al. (eds.) *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*, pp.27-46. Amsterdam: John Benjamins.
- 投野由紀夫. (2007). 『日本人中高生一万人の英語コーパス: JEFLL Corpus: 中高生が書く英文の実態とその分析』 小学館.
- 投野由紀夫, 金子朝子, & 杉浦正利. (2013). 『英語学習者コーパス活用ハンドブック』 大修館書店.

研究論文における英語と日本語のアブストラクトは読み手を意識しているか

浅野 元子(大阪医科薬科大学)*

松田 紀子(近畿大学)

*motoko.asano@ompu.ac.jp

Are English and Japanese abstracts in a research paper written for the audiences?

ASANO Motoko (Osaka Medical and Pharmaceutical
University)
MATSUDA Noriko (Kindai University)

Abstract

本研究は, 論文における英語と日本語のアブストラクトが各々の読み手を意識した言語的特徴を有するかどうかを探索した。2020 年 1 年分の教育心理学研究誌 (EP) 30 報と日本公衆衛生雑誌 (JPH) 30 報の英文と日本語のアブストラクトをコーパス化し, 修辞の運び (Move-step) を分析した。EP と JPH の総語数は英文が約 6,000 語と 11,000 語 (CasualConc), 日本語は約 7,500 と 14,000 語 (WinCha および CasualConc) であった。EP の 12 編 (40.00%), JPH の 17 編 (56.67%) は英文と日本語の論理の運びに差異があり, EP は英文または日本語どちらかのみで生起する Move-step によって詳述することがあり, 特に研究の手順を英文で詳述する傾向がみられた。JPH は, 論理の運びに違いがみられ, 副次的な結果・問題点などの記述が日本語に多い傾向がみられたが, 抄録全体では英語と日本語で類似した情報が提示されるように見受けられた。各々の読み手を意識して調整されていることが示唆された。

Research articles published in Japanese journals tend to have English and Japanese abstracts. This study investigated the linguistic features of English and Japanese abstracts written for articles published in two Japanese journals: the Journal of Educational Psychology (EP) and the Japanese Journal of Public Health (JPH). Corpora were constructed using English and Japanese abstracts of 30 articles each from the two journals. The EP and JPH corpora had approximately 6,000 and 11,000 words in English (CasualConc) and 7,500 and 14,000 words in Japanese (WinCha and CasualConc), respectively. Swalesian move-step analysis revealed the differences in the occurrence of move-steps between the English and Japanese texts in 12 (40.00%) and 17 (56.67%) of the EP and JPH articles, respectively. The findings suggested that the descriptions in English and Japanese texts are fine-tuned for the audiences.

Keywords

抄録, abstract; ジャンル, genre; move, ムーブ; corpus, コーパス

1. はじめに

ジャンルは、目的を持って行われるコミュニケーションの種類といわれており (Hyon, 2018), 情報の受け手に応じて形作られることが知られて久しい (Swales, 1990)。研究論文のアブストラクト (以下、抄録とする) は、“part-genre” (Swales, 2004, p. 239) の一つとされ、論文テキストを代表するもの、結晶、要旨などさまざまな称されてきた (Hyland, 2000)。近年、同一ジャンルのテキストでも情報の受け手が基盤とするコミュニティの規範に沿うことが望ましいとされる (Connor, 2011)。本研究は、同一論文に掲載された英語と日本語の抄録が、各々の読み手を意識して書かれているかどうかを検討する目的で、テキストの言語的特徴を調査した。

2. 先行研究

2.1 ムーブ分析

ムーブ分析は、Swales (1990) により提案されたジャンル分析の枠組みの中核をなすテキスト分析手法で、Swales (1981) を端緒とする。研究論文の緒言部において書き手の意図を伝えるための修辞の運びが特定され、move (ムーブ) と称された。ムーブは「コミュニケーション上の役割を担う談話または修辞の単位」(Swales, 2004, p. 228) である。ムーブの中には、「ともに、または組み合わせるムーブを示す」ための step (ステップ) と呼ばれるテキストがあり、ステップは主に「ムーブの目的を果たすために機能する」(Biber, Connor, & Upton, 2007, p. 24)。「Move-step 分析」(Swales, 1990, p. 142) において、ステップレベルの解釈は、ムーブの解釈より具体的である。

2.2 学術テキストの対照研究

テキストの文化的背景に注目した対照研究によって、テキスト構成は言語文化によって異なることが示されてきた (Kaplan, 1966; Hinds, 1983; 澤田, 1978)。しかし近年、テキスト構成には発信者の居住地域の文化に加えて学術、教室などの文化が複雑に関与すると説かれる (Connor, 2004)。なかでも、学術テキストは、「ライティングの文化」「writing culture」の影響を受けるが (Moreno & Swales, 2018, p. 50), 「型と変種の両面で共通語の機能を持つ伝達手段」となる傾向が強まっている (Swales, 2017, p. 251)。学術論文テキストにおける母語の影響を調査するには「ステップの修辞的な多様性を特定する」ことが有用とされる (Moreno & Swales, 2018, p. 44)。

2.3 抄録の言語的特徴に関する研究

抄録は論文の本文とは別に「論文全体を考慮して描出」されることが知られる (Bazerman, 1984, p. 58)。医学論文の抄録を調査した Salager-Meyer (1990) は、目的、方法、結果、結論という4つのムーブで構成されると読みやすく、ムーブの運びが逸脱すると「論旨が伝わりにくくなる」と指摘している (p. 370)。抄録800編を専門家の協力を得て分析した Hyland の研究 (2000) で

は、先行研究 (Salager-Meyer, 1990) による指摘にもかかわらず、背景、目的、方法、結果、結論というムーブの運びを実現しているものが少ないと指摘される。

3. リサーチデザイン

3.1 研究目的と研究設問

本研究の目的は、日本語論文における英語と日本語の抄録がそれぞれの読み手を意識した言語的特徴を有するかどうかを探索することである。以下のリサーチクエスション (RQ) を設定した。

RQ1 英語と日本語の抄録における各々の Move-step の生起率は同じであろうか。

RQ2 英語と日本語の抄録における Move-step の運びは同じであろうか。

RQ3 英語と日本語の抄録に特徴的な言語使用の例はあるであろうか。

3.2 コーパス

2020 年 1 年分の教育心理学研究誌 (EP) 30 報の英語と日本語の抄録、および同年の日本公衆衛生雑誌 (JPH) 30 報の英文と日本語の抄録を用いてコーパスを構築した。コーパス編纂に際しては “representativity, reputation, and accessibility” を考慮するという Nwogu (1997, p. 121) の考えに基づいて、日本国内誌を対象とする Google Scholar Metrics などを考慮した。

3.3 手法

3.3.1 データの事前処理

EP, JPH ともに、ウェブサイト上の日本語のタイトルと抄録をコピーして用いた。英語に関しては、論文の PDF ファイルは、大半が読み取り専用のため、「Google レンズ」を使用または印刷物をスキャンして OCR 処理し、文字化け等を目視で確認しながらスプレッドシートに入力した。なお、EP は構造化された抄録はなく、JPH は全編が目的、方法、結果、結論に構造化されていた。

3.3.2 分析の手順

抄録は、発表者の1名が EP, JPH の順、もう1名が JPH, EP の順でテキストの修辭的な役割 (Hyland, 2000; Salager-Meyer, 1990) に注目して Move-step に分類し、一致率 (Rau & Shih, 2021) を算出した。一致率は、EP の英文が 96.19%, 日本文が 95.59%, JPH の英文が 77.31%, 日本文が 85.88% であった。Move-step は Zoom 会議で画面共有して話し合いにより最終化した。

3.3.3 Move-step コーパスの構築と分析

コーパスから Move-step テキストを抽出した。日本文は、奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 (松本研究室) の形態素解析器 WinCha を用いて形態素に分けた (石田・小林, 2013)。Move-step テキストは、CasualConc (Imao, 2021) を用いて計量した。

4. 結果と考察

4.1 RQ1 英語と日本語の抄録における Move-step の生起率

テキストにみられた Move-step を図 1 に、生起率を図 2 に示した。

EP は Move-step 13 が1編 (EP025) を除いてすべての抄録に生起し、必須と考えられた。Move-step 21 についても英日ともに 8 割以上、Move-step 22 についても英文 8 割以上、日本語 7 割以上、Move-step 32 については英日ともに 7 割以上の抄録に出現した。反対に Move-step 12b, 34, 42, 44, 45 の生起率は 1 割以下であった。42 は英文 1 編のみに出現したが、44 は英日ともに認められなかった。Move-step 22 は英文では 25 編 (83.33%)、日本文では 22 編 (73.33%) に生起し、英文に多く出現する傾向がみられた。

図1 アブストラクトに観察されたMove-step

Background	11	研究の背景情報を提示
	12	関連する研究のレビュー
	12b	問題点 (リサーチギャップ)
Purpose	13	この研究の紹介・目的
Methods	21	研究のデザイン・対象・データ収集方法
	22	研究の手順
	23	分析の手順
Results	31	集まったデータの背景情報
	32	主な結果
	33	副次的な結果・問題点など
	34	結果の解釈
Conclusions	41	主な結果とその意義
	42	背景・先行研究に照らした考察
	43	今後の展望 (さらなる研究)
	44	この研究の限界
	45	考察内容の提示
	50	判定不能

JPH は、Move-step 13, 21, 32, 41 が1つ (PH023 における 13) を除いてすべての抄録に生起し、必須と考えられた (図 2)。Move-step 23 についても英日ともに 8 割以上の抄録に出現した。反対に Move-step 34, 42, 44, 45 の生起率は 1 割以下であった。44 は英文 1 編、45 は日本文 1 編のみに出現したが、Move-step 33 は英文では 14 編 (46.67%)、日本文では 17 編 (56.67%) に生起し、日本文に多く出現する傾向がみられた。

図2 Move-stepの生起率

Move-step	11	12	12b	13	21	22	23	31	32	33	34	41	42	43	44	45	50	
EP	EN	16.67	13.33	3.33	96.67	86.67	83.33	43.33	26.67	73.33	20.00	6.67	56.67	3.33	13.33	0.00	13.33	6.67
	JA	16.67	13.33	3.33	96.67	83.33	73.33	43.33	30.00	73.33	20.00	6.67	63.33	0.00	13.33	0.00	10.00	6.67
JPH	EN	43.33	36.67	26.67	96.67	100.00	36.67	83.33	73.33	100.00	46.67	10.00	100.00	3.33	50.00	3.33	0.00	0.00
	JA	46.67	33.33	26.67	100.00	100.00	40.00	86.67	70.00	100.00	56.67	6.67	100.00	3.33	56.67	0.00	3.33	0.00

EN: 英語抄録, JA: 日本語抄録

4.2 RQ2 英語と日本語の抄録における Move-step の運び

EP では、12 編 (40.00%) において Move-step の運びに違いがみられた。英文か日本文のどちらかのみが生起する Move-step の存在が特徴的であり、研究の目的 (Move-step 13) に関する記載を英文のみに叙述する例として、“The performance of the proposed model was evaluated both in a simulation study and when applied to data from an actual math achievement test.” (EP028) があげられる。また、研究の主な結果を日本文のみに叙述する例として、「以上より、因子構造の安定性に課題は残るものの、一定程度の妥当性と信頼性が確認された。」 (EP013) があげられる。

JPH では、17 編 (56.67%) において Move-step の運びに違いがみられた。「健康日本 21 (第二次) の目標を達成するため、各自治体は健康課題を適切に評価し、保健事業の改善につなげることを求められている。」という日本の状況を日本文のみに叙述する例 (JPH002)、“Odds ratios (ORs) and . . . were estimated” という分析方法を英文では方法、日本文では結果セクシ

ョンの中ほどに「... オッズ比を比較した結果, ...」と記載する例 (JPH006) などがみられた。

4.3 RQ3 英語と日本語の抄録に特徴的な言語使用の例

EP では, APA (American Psychological Association) style を使用して叙述しているためか, 英文抄録には *we* は出現せず, 20 編に *the (The) present study* が 23 (内 21 件は研究の目的を述べる Move-step) で出現した。その対訳と考えられる「本研究」は 27 編に 34 件出現し, その内の 18 件が *the present study* の対訳となっていた。

JPH では, 英文抄録 19 編に *we* が 49 件出現し, 医学論文の言語特徴を調査した先行研究 (Nwogu, 1997; Salager-Meyer, 1992) の例文と類似した特徴が示されたが, 読み手を巻き込んだ *we* (Hyland, 2001) は 1 件 (JPH019) のみであった。英文にみられた *we* の対訳は, 「本研究は」または「本研究では」が 6 件, 「我々は」が 1 件, 「筆者らは」が 1 件で, 対訳の出現は 1 編あたり 1 件のみであった。また, 日本文では「応急仮設住宅」, 「要支援高齢者」, 「食品群別摂取量」のような漢字による専門語彙が反復して出現し, 対応する英文では多様な言い換えがみられた。

5. まとめ

テキストの言語的特徴を分析した結果, EP における同一論文に掲載された英語と日本語の抄録には, 英文または日本文どちらかのみで生起する Move-step が散見され, 特に研究の手順を英文で詳述する傾向がみられた。Move-step の数には差があり, Move-step の生起率や運びには違いが認められた。著者による多少の個人差はあると考えられるが, 各々の読み手を意識して書かれている可能性があるといえる。JPH は, 過半数において Move-step の運びに違いがみられ, 副次的な結果・問題点などの記述が日本文に多い傾向であったが, 各々の抄録全体でみると英語と日本語で類似した情報が提示されるように見受けた。言語によって情報の提示の仕方を工夫していると考えられた。専門語彙を日本語では反復して, 英語では言い換えて記述する傾向がみられた。英語抄録における *we* は日本語では「本研究は」などの対訳が 1 編あたり 1 件出現するのみであった。いずれの学術誌も, 一つの研究について英語と日本語で抄録にまとめる際に, 各々の言語の読み手を意識して情報の示し方を調整していることが示唆された。

引用文献

- Bazerman, C. (1984). The writing of scientific non-fiction: Contents, choices, constraints. *PRE/TEXT* 5: 39–74.
- Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins.
- Connor, U. (2004). Intercultural rhetoric research: Beyond texts. *English for Specific Purposes*, 3(4), 291–304.
- Connor, U. (2011). *Intercultural rhetoric in the writing classroom*. The University of Michigan Press.

- Hinds, J. (1987). Reader versus writer responsibility: A new typology. In U. Connor & R. B. Kaplan (Eds.), *Writing across languages: Analysis of L2 text* (pp. 141–152). Addison-Wesley Publishing Company.
- Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. The University of Michigan Press.
- Hyland, K. (2001). Bringing the reader. *Written communication*, 18(4), 549–574.
- Hyon, S. (2018). *Introducing genre and English for specific purposes*. Routledge.
- Imao, Y. (2021). CasualConc (Version 2.1.7) [Computer software]. Retrieved from <https://sites.google.com/site/casualconcj/>
- Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning*, 16, 1–20.
- Moreno, A. I. (1997). Genre constraints across languages: Causal metatext in Spanish and English RAs. *English for Specific Purposes*, 16(3), 161–179.
- Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, 50, 40–63.
- Nwogu, K. (1997). The medical research paper: Structure and function. *English for specific purposes*, 16(2), 119–138.
- Rau, G., & Shih, Y. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *English for Academic Purposes*, 53, 101026.
- Salager-Meyer, F. (1990a). Discoursal flaws in medical English abstracts: A genre analysis per research- and text-type. *Interdisciplinary Journal for the Study of Discourse*, 10(4), 365–384.
- Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes*, 11(2), 93–113.
- Swales, J. M. (1981). *Aspects of article introductions*. The Language Studies Unit, Aston University.
- Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.
- Swales, J. M. (2017). Standardisation and its discontents. In M. Cargill & S. Burgess (eds). *Publishing research in English as an additional language: Practices, pathways and potentials*. The University of Adelaide Press.
- 石田基弘・小林雄一郎 (2013) 『R で学ぶ日本語テキストマイニング』 ひつじ書房.
- 澤田治美 (1978) 「日英語文副詞類 (Sentence Adverbials) の対照言語学的研究—Speech act 理論の視点から—」 『言語研究』, 74, 1–36.

Development of the Ballet English Corpus (Ver. 1.0) for Art Major Students

USAMI Hiroko (Tokai University)

hiroko.usami@tokai-u.jp

Abstract

This study aims to describe a specialised corpus, the Ballet English Corpus (BEC Ver. 1.0), which contains approximately 1 million words that make up written texts in 10 different categories related to classical ballet that adults and children are expected to use. This study provides a description of the corpus design and analysis of the word lists and keyword list in the BEC. The analysis of the BEC indicates that the category for ‘people’, by far, contains the most data. In addition, the word *ballet* frequently appears, as well as articles, a conjunction, prepositions and *be* verbs. Furthermore, the word *ballet*, the lemma DANCE, French words, repertoire names, role names, and dancers’ ranks occur as keywords.

Keywords

specialised corpus, ESP for art, classical ballet, word list, keyword list

1. Introduction

English for specific purposes (ESP) analyses English used in various fields, and specialised corpora containing specific fields of English, for example, analyse frequent vocabulary and grammar. However, studies examining the use of English in classical ballet remain limited, although classical ballet has been enjoyed for different purposes by people of all ages in Japan. This study aims to describe a specialised corpus created by the author, the Ballet English Corpus (BEC), which can be applied to the study of ESP. The BEC (Ver. 1.0) currently contains approximately 1 million words of written texts used in 10 different contexts related to classical ballet (ballet techniques, companies, studios, history, schools, theatres, people, narratives, repertoires, and miscellaneous) that both adults and children are expected to use. This study provides a description of the corpus design and analysis of the word lists and keyword list in the BEC.

2. Literature Review

2.1 Textbooks and Specialised Corpora for ESP

Among the ESP textbooks, which have been published by 11 Japanese and overseas

publishers for university English textbooks and are available in Japan in 2022, the popular genres of ESP textbooks are science, medicine (including nursing and dentistry), and business. Actually, some ESP textbooks for art and music and a lot of ESP textbooks for movie have been published. However, most of the ESP textbooks for music teach English in restricted contexts; how to understand song lyrics.

Besides the ESP textbooks strictly available for university teachers and students, a couple of ESP textbooks for music were published and can be accessed by anyone involved in or is interested in music, such as music majors in high school and university (e.g., Kubota & Orui, 2017). These textbooks help those interested in learning English conversation in music learning contexts, such as through music lessons, accompanying musical instruments, and concerts. For dance and theatre studies, especially classical ballet, only a few ESP textbooks for dance have been published. In particular, only one textbook for learning English in classical ballet situations has recently been published for children interested in learning classical ballet overseas. The textbook aims to help them learn English they can use in various contexts related to classical ballet, such as when attending classes, watching performances, shopping for dancewear in ballet stores, and participating in school activities (Ito, 2021).

2.2 Popularity of Classical Ballet in Japan

Classical ballet has been especially popular in Japan across age groups. Approximately 256,000 people learned classical ballet in Japan in 2021, and more than 70% of classical ballet studios teach children in elementary school or younger and adults in their 40s and 50s. (Showa Academia Musicae Ballet Research, 2022). Moreover, approximately 13% of the classical ballet studios have students who belong to overseas classical ballet schools and companies. They start their career as professional dancers in their late teens or early twenties (Oyama & Umino, 2022, Showa Academia Musicae Ballet Research, 2022). Actually, the number of Japanese classical ballet dancers working in overseas classical ballet companies has gradually been increasing from the 2017 to 2018 season to the 2021 to 2022 season. Regardless of age, most classical ballet learners enjoy classical ballet by watching performances and taking lessons. However, only a few classical ballet teachers have learned or are learning how to teach classical ballet formally, and many do not seem to conduct research concerning classical ballet education (Oyama & Umino, 2022; Showa Academia Musicae Ballet Research, 2022).

3. Ballet English Corpus (Ver. 1.0)

Based on the situation of ESP for classical ballet, the popularity of classical ballet in

Japan, and the increasing number of Japanese classical ballet dancers in overseas classical ballet companies, the author conducted a specialised corpus—the BEC—so that English used in the classical ballet contexts can be examined and analysed. Based on the book *Ballet: The Definitive Illustrated Story* (Durante, 2018), written texts from books and websites based on the 10 categories were collected: ballet basics, company, dance studio, history, narrative, people, repertoire, school, theatre, and miscellaneous. All the words in the books except for their contents, index, glossary, and photos were manually typed into text files.

By far, the category ‘people’ has the largest number of words and accounts for almost half of the total. Most of the words were obtained from websites for adults introducing classical ballet dancers. The second and third largest categories are ‘repertoire’ and ‘company’. Most words were obtained from books for children and websites for adults, respectively. The categories ‘ballet basics’, ‘school’, and ‘miscellaneous’ are small, and the categories ‘history’ and ‘theatre’ are even smaller. While the ratio of each category is biased, the result is a realistic representation.

At the moment, the BEC (Ver. 1.0) currently contains approximately 1 million words from written texts used in classical ballet situations. In the future, the author plans to expand the BEC by adding more books and websites, which can be bought and are available online. In addition, more books on classical ballet should be made available overseas, as well as pamphlets and leaflets in theatres, dance schools, or dance studios, which can also be made available for purchase by those buying tickets to the show.

The software package, #LancsBox, which was developed at Lancaster University, was used to analyse the BEC (Ver. 1.0). The Words (for wordlists), Ngrams (for n-grams), and KWIC (for concordances) functions available in #LancsBox were used in this research.

Using the BEC (Ver. 1.0), the following research questions are addressed in the following sections:

RQ1: What words frequently occur in the BEC (Ver. 1.0)?

RQ2: Which 2-, 3-, and 4-grams frequently occur in the BEC (Ver. 1.0)?

RQ3: What are the keywords found in the BEC (Ver. 1.0)?

4. Results and Discussions

4.1 Word List in the BEC

First, in order to examine which single word is frequently used in the BEC (Ver. 1.0), a word list was obtained. According to the word list, articles (*the*, *a*, and *an*), a conjunction (*and*), prepositions (*in*, *of*, *to*, *with*, *for*, *as*, *at*, *on*, *by*, and *from*), and *be* verbs (*was*, *is*, *be*, and *are*) frequently occur in the BEC. In addition, second- and third-person

pronouns such as *you*, *your*, *her*, *she*, *he*, and *his*, with slightly more third-person pronouns of female (*her* and *she*), appear more frequently than the first-person pronoun *I*. Of course, words related to classical ballet occur; *ballet* (rank 7), *dance* (rank 18), and *dancer* (rank 27). Besides English words, the French word *de* also appears. Words related to ballet companies or schools were also found; *royal*, *school*, and *company*.

4.2 N-grams in the BEC

Next, in order to examine what phrases occur in the BEC (Ver. 1.0), 2-, 3-, and 4-grams were obtained. In the 2-gram list, phrases composed of a preposition/conjunction + article occur frequently (*of the*, *in the*, *at the*, *to the*, *and the*, *with the*, *for the*, *as a*, and *on the*). The word *ballet*, which also occurs quite frequently in the word list, appears in 2-, 3-, and 4-grams and collocates with other words as the names or part of the names of ballet companies or schools (*royal ballet*, *ballet school*, *national ballet*, *the royal ballet*, *english national ballet*, *royal ballet school*, *the royal ballet school*, *new york city ballet*, *of the royal ballet*, *at the royal ballet*, *the national ballet of*, *national ballet of canada*, and *school of american ballet*), as well as dancers' ranks (*corps de ballet*, *the corps de ballet*, *corps de ballet in*, *dancer corps de ballet*, and *ballet dancer corps de*). Moreover, the words *dance* and *dancer* appear frequently in the word list but do not frequently collocate with other words. Furthermore, second- and third-person pronouns (*you*, *your*, *her*, *she*, *he* and *his*) show frequent occurrence in the word list but do not directly and frequently collocate with other words nor appear in 2-, 3-, or 4-grams. The French word *de* is used with other words as the names or part of the names of steps (*pas de* and *pas de deux*) and dancers' ranks (*corps de ballet*, *the corps de ballet*, *corps de ballet in*, *dancer corps de ballet*, *of the corps de*, and *ballet dancer corps de*).

Although the word *age* does not appear in the single word list or the 2-gram list, phrases indicating age occur in 3- and 4-grams (*the age of*, *at the age*, and *at the age of*), especially in the 'people' category. Similarly, although the word *member* does not appear in the single word list, nor in 2- or 3-grams, phrases indicating membership occur only in 4-grams (*a member of the*, *member of the corps*, and *as a member of*), especially in the 'people' category. Furthermore, phrases indicating ticket information occur only in 4-grams (*more tickets & info* and *tickets & info july*), especially in the 'people' category, specifically dancers' biographies in one particular ballet company. However, the word *member* does not appear in the single word list, nor in 2- or 3-grams.

Phrases more specific to ballet used as the names or part of the names of the repertoire occur in 2- and 3-grams (*the nutcracker*, *swan lake*, *the sleeping beauty*, and *romeo and juliet*). In addition, the phrase *was born in* occurs in 3-grams, followed by the place

dancers or staff were born in. Similarly, the phrase *was promoted to* occurs in 3-grams followed by the dancers' rank, such as principle, soloist, or artist. One more characteristic phrase, *under the direction of*, occurs, followed by teachers, mothers, directors, or choreographers, in 4-grams.

4.3 Keyword List in the BEC

Finally, in order to examine which single word is characteristically used in the BEC, the keyword list was obtained, comparing the word list of the BEC with that of the BNC2014-baby containing 5,024,072 tokens stored in the #LancsBox.

Table 1 *Keyword list in the BEC (Ver. 1.0)*

Rank	Word	Stats.	Rank	Word	Stats.	Rank	Word	Stats.	Rank	Word	Stats.
1	ballet	160.87	11	pas	13.79	21	written	11.32	31	artist	9.58
2	dance	34.71	12	prince	13.19	22	romeo	11.17	32	dancing	9.50
3	dancer	34.40	13	giselle	13.16	23	juliet	11.17	33	repertoire	9.37
4	royal	19.32	14	corps	13.11	24	fairy	10.97	34	theatre	9.34
5	nutcracker	19.20	15	roles	12.19	25	ballets	10.74	35	la	9.25
6	de	16.68	16	website	12.18	26	beauty	10.56	36	deux	8.87
7	principal	15.13	17	lake	11.97	27	danced	10.16	37	classical	8.78
8	dancers	15.07	18	opera	11.90	28	performed	9.93	38	role	8.49
9	swan	14.89	19	company	11.89	29	sleeping	9.83	39	featured	8.30
10	soloist	14.35	20	joined	11.42	30	cinderella	9.80	40	school	8.27

The word *ballet*, by far, appears the most frequently, approximately five times more than the 2nd word in the rank. Compared with the single word list and 2-, 3-, and 4-grams, the lemma DANCE (*dance*, *dancer*, *dancers*, *danced*, and *dancing*) appears, which indicates that it is a keyword used as different parts of speech and verb forms. Similarly, more varied words related to ballet companies or schools (*royal*, *opera*, *company*, *theatre*, and *school*) and more French words (*de*, *pas*, *corps*, *la*, and *deux*) occur as keywords, although some of the keywords do not appear in the single word list or 2-, 3-, and 4-grams. Furthermore, various words used as the name or part of the names of a repertoire occur as keywords; *nutcracker* in *The Nutcracker*, *swan* and *lake* in *Swan Lake*, *giselle* in *Giselle*, *romeo* and *juliet* in *Romeo and Juliet*, *beauty* and *sleeping* in *The Sleeping Beauty*, and *cinderella* in *Cinderella*.

Surprisingly, words that do not occur in the single word list or 2-, 3-, and 4-grams occur as keywords; those related to roles or characters classical ballet dancers portray (*prince*,

roles, and *fairy*) and those related to the rank of classical ballet dancers in the company (*principal*, *soloist*, and *artist*). Finally, verbs such as, *joined*, *performed*, and *featured*, occur characteristically as keywords, although the verb phrase *was born in* and *was promoted to* occur in 3-grams.

5. Conclusion

This paper described the corpus design and analysis of a specialised corpus, the BEC. Reflecting classical ballet situations in Japan, the author constructed the BEC to examine the English used in classical ballet situations. Currently, the BEC (Ver. 1.0) contains approximately 1 million words grouped into 10 categories with the majority of words found under ‘people’.

In future research, the BEC (Ver. 1.0) can be expanded by adding more data to each category or including more categories. Once the BEC contains enough data, it can be used to teach English to art majors. Characteristically, the BEC contains texts for both adults and children. Therefore, the texts should be analysed separately for adults and children for different purposes. In addition, the BEC consists of 10 different categories and each category should have specific features. Words or phrases that are frequently used in each category can be identified for different purposes. Based on analyses of words and keyword lists, vocabulary lists and textbooks can be compiled for children who learn classical ballet and those who intend to study abroad, as well as professional dancers working at overseas ballet companies.

Bibliography

- Durante, V. (2018). *Ballet: The Definitive Illustrated Story*. DK.
- Ito, Y. (2021). *Ballet eikaiwa*. [Ballet English conversation]. Shinsyokan.
- Kubota, K. & Orui, T. (2017). *Kaitei ban. Ondaisei ongakuka no tamen oigo de step up—ongaku ryugaku de yakudastu eikaiwa 50 scene*. [Step up in English for music major students and musicians—English conversations in 50 scenes that are helpful in music study abroad—revised edition]. Stylenote.
- Oyama, K. & Umino, B. (2022). *Nihon no ballet kyoiku kannkyo no jittai bunseki. Ballet kyoiku ni kansuru zenkoku chosa 2021—kihon houkoku*. [Analysis of the ballet education situation in Japan: Nationwide survey on ballet education—basic report]. Showa Academia Musicae.
- Showa Academia Musicae Ballet Research. (2022). *Nihon no ballet kyoiku ni kansuru zenkoku chosa—hokokusyo*. [Nationwide survey on ballet education in Japan—report]. Showa Academia Musicae.

COCA における正誤表現と非標準表現について

田畑 圭介(神戸親和女子大学)

tabata@kobe-shinwa.ac.jp

On (In)correct and Non-Standard Expressions in COCA

TABATA Keisuke (Kobe Shinwa Women's University)

Abstract

This paper defines criteria for determining how unconfirmed English expressions that can be distinguished from incorrect ones can be specified as non-standard or standard usages. Focusing on phrases such as “got me to doing,” “got me done” and “all you want,” the following hypotheses will be derived. (1) If the proportion of the frequency of an expression to that of an corresponding correct expression is 15% or more in COCA and its certain uses are attested in the TV/Movie section, it is classified as non-standard (but frequent) expression. (2) If an expression is approximately as frequent and productive as the corresponding correct expression in COCA, or if instances of an expression that does not have a corresponding correct expression are detected in 200 or more cases in COCA, it is identified as standard (=correct) expression.

Keywords

COCA, 非標準表現, 辞書学

1. はじめに

COCA (Corpus of Contemporary American English) は現在の英語研究のデータ資料として広範囲に使用されているコーパスの一つである。約 10 億語からなる COCA は、8 つのジャンルで構成され、現代アメリカ英語の研究調査を支える言語資料となっている。Davies (2015:26) は COCA について、しっかりとした (robust)、バランスのとれた (well-balanced) コーパスと表現している。Davies (2015) で紹介されている COCA は、Spoken, Fiction, Magazine, Newspaper, Academic の 5 つのジャンルで構成されていたが、2022 年 5 月現在の COCA には新たに Web (General), Web (Blog), TV/Movies の 3 つのジャンルが追加されている。

英語の用法を調べる際、英語の正用法が調査対象となるのが一般的であり、その調査データは正用法のみで構成されるのが自然である。例えば、英語の各種辞典は正用法を確認する資料として最適であり、英語学習上、重要な参照資料となっている。

COCA を現代アメリカ英語の調査資料として用いた場合、10 億語を超えるデータが用法調査の裏付材料として活用されることになる。ただし、COCA でその使用が確認できたとき、それが標準用

法として認定できるかどうかは慎重を要する。例えば「～するようになる」という英語表現は、become to do は誤用法で、come to do などが正用法となるが、COCA では実際のところ、誤用とされる become to do の使用例が見つかる。

表1 COCA における become/come to do の頻度

	become	come
～ to believe	2 (0.12%)	1730 (99.88%)
～ to know	3 (0.12%)	2420 (99.88%)

[become] to believe は 2 例見つかり、[become] to know は 3 例検出されるが、正用法となる [come] to believe は 1730 例、[come] to know は 2420 例が検出されるので、誤用とされる become to do の使用比率は [become] to believe が 0.12%、[become] to know も 0.12% である。実際の使用例が COCA で検出されたとしても、become to do のように誤用とされる例も見つかることから、COCA を使用する際にはあらかじめ COCA が正用法だけで構成されるコーパスではない点をおさえておく必要がある。

本稿では、文法規則から逸脱し、誤用とされる英語定型表現が COCA に一定数含まれている現状を踏まえつつ、誤用例の頻度を大きく上回って使用されている非標準用法の英語表現に着目する。正用法ではないことから、英語学習者はその使用は避けるべきと帰結されるが、日常的にその使用が散見される表現形式を英語学習者に補足情報とともに伝達することは、英語学習上有用だと考えられる。英語学習者が正用法ではない英語表現に出会い、その解釈が求められる場面が時に起こりうるためである。正用法以外の表現が含まれる COCA を検索対象とし、英和辞典や英語文法書への非標準的用法の掲載の可否を判定する数値基準について、2 節より考察を行う。また新規に正用法と判断する際の判定基準についても合わせて検討する。

2. get A to do/to doing 構文

get A に続く動詞形は to do が正しく、to doing は認められない。COCA では get A to do の A に入る要素は代名詞が多く、後続する動詞は do, talk, stop, come, go, say などである。get A to do は動名詞を後続させることがないはずだが、COCA で [get] me to *を検索すると、got me to thinking が 150 例検出され、3 位の get me to go の 2 倍ほどの頻度でその使用例が見つかる。(1) のように got me to thinking は「考え始めるようになった」という意味で用いられていることから、この表現は got me to think からの派生ではなく、got me thinking が元表現になると推察できる。

(1) I guess that got me to thinking about the widow Audel.

そのことで私は未亡人のオデルさんに思いを寄せるようになったと思います。

Chocolat『ショコラ』(2000) <00:42:50>

表2 COCA における got me thinking / to thinking の頻度

	got me thinking	got me to thinking
--	-----------------	--------------------

頻度	800 (84.21%)	150 (15.79 %)
----	--------------	---------------

got me to thinking は、主な英和辞典や語法書において関連の記述は確認できない。それでも COCA で 150 例 (BLOG:50, WEB:40, TV/M:20, SPOK:14, FIC:11, MAG:2, NEWS:8, ACAD:5) が検出され、正用法に対して 15% という比率でその使用例が見つかり、また入念なチェックを経て作り出される TV/Movie のスクリプトで 20 例の使用が確認できることから、非標準の表現として認定することは妥当だと考えられる。TV/Movie の「話し言葉コーパスよりもより口語的 (colloquial) で、より中核的 (core)」 (https://www.english-corpora.org/coca/help/coca2020_overview.pdf) とされる特徴から、現代の非標準的表現を検出するデータ資料として TV/Movie のセクションは中核的な役割を果たすと推測でき、正用法との比率を加味しつつ、次の仮説を導くことにしたい。

(2) 仮説1 正用法との比率が 15% を超え、TV/Movie で一定の使用が確認できる用法は、非標準用法としてその存在を認定できる。

COCA はアメリカ英語のコーパスであることから、イギリス英語のコーパス等も合わせて検索し、検出状況によっては((米))あるいは((主に米))といったレーベル付与が必要となる。

3. get A done 構文

get A 動詞の構文では get A to do, get A doing に加えて、get A done の構文が存在する。get A done の各表現は、英和辞典では[使役]、[完了]、[被害]の 3 つのタイプに分類されている。got me killed/confused/kicked/elected/mixed/pegged/laid はそれぞれ「～させる」という意味には結びつくが、厳密には「私が～されるようにさせた」という表現形式となっており、結果的に「私を～した」という解釈になる。got me killed は[使役][完了][被害]の 3 分類では[使役]に属するが、それぞれの意味の差異を考慮すると、別の新たな区分への分類が妥当だと考えられる。

表 3 COCA における get me done 構文

使役		行為	
GOT ME STARTED	93	GOT ME KILLED	71
GOT ME INTERESTED	86	GOT ME CONFUSED	33
GOT ME HOOKED	68	GOT ME KICKED	22
GOT ME FIRED	57	GOT ME ELECTED	19
GOT ME EXCITED	48	GOT ME MIXED	18
GOT ME WORRIED	47	GOT ME PEGGED	15
GOT ME INVOLVED	33	GOT ME LAID	14

表 3 右列の表現群は「私を～させた」ではなく、「私を～した」という解釈になる。get A done の[使役][完了][被害]の 3 分類とは別に、[行為]という新区分を立て、次のように記述することが可能である。

(3) get A done

『行為』 <人が>A<人>を...する(主語は行為者で、A は被害者・対象者を表す)

“You got that man killed.” “I got him killed?” 「君があの人を殺したんだ」「俺が殺した？」

He's got me confused with someone else. 彼は俺のことを別の人と勘違いしている。

「行為」に分類される **got me killed** は表 3 が示すように、**got me started** などの純粋な[使役]のグループと同程度の頻度で用いられており、また右列の表現群には生産性が見られる。[行為]の **got me done** 構文については、次の帰結が妥当である。

(4) 仮説 2: COCA において正用法と同程度の頻度と生産性が確認できる用法は、正用法としてその存在を認定できる。

文法的で生産性が見られる表現であることから、正用法として取り扱うことができる。ただし、英和辞典の従来の記述枠には当てはまらないため、新たな項目で取り扱うべき表現群となる。

4. 比較対象のない表現例: **all you want**

大橋(2013)は、名詞用法から副詞用法への文法化で生まれた **all you want** (好きなだけ) を認知言語学の観点で分析している。

(5) a. You can use it all you want.

b. You can inspect it all you want, but it looks perfect to me. 大橋(2013:166)

この副詞用法の **all you want** はジーニアス英和大辞典に記述が見られる英語表現である。

(6) **all one's wants** [副] (略式) 心ゆくまで (to one's heart's content)

You can daydream all you want. 君は思う存分空想にふけることができるよ。

COCA で文末の **all you want** を検索すると、625 例が検出され、その多くが会話体で用いられている。これは **ready when you are** (232 例)、**I couldn't agree with you more** (220 例)、**Not that I know of** (462 例) といった他の慣用的な会話表現よりも高い頻度を示している。副詞用法は名詞用法から転じたものだが、副詞用法と名詞用法は品詞が異なるため、名詞用法との比率を副詞用法の認定基準に用いることはできない。この場合、COCA での純粋な頻度で判定することになる。副詞用法の **all you want** については次の基準を設けることにしたい。

(7) 仮説 3: COCA において 200 以上の用例が確認され、正用法との対立がない表現について

は、正用法と認定できる。

5. COCA における数値基準

仮説 3 は、COCA において 200 例以上という基準をもとに判定しているが、この 200 という数値を検証する一例として、**absolutely** が形容詞に求める選択制限を活用してみる。小西 (編) (1989) の **absolutely** の項目では、形容詞を修飾する場合には、その意味尺度の両極端にくるもののみに修飾可能とし、*That is absolutely good [bad]. / That is absolutely wonderful [terrible]. の例が提示されている。それぞれの頻度調査の結果を表 4 に示す。

表 4 COCA における absolutely +形容詞の頻度

	* good	wonderful	* bad	terrible
absolutely ~	35	318	12	148

容認されない **absolutely good** や **absolutely bad** はそれぞれ 35 例、12 例と低頻度となっており、200 の数値には及ばない結果となっている。一方で正用法の **absolutely terrible** は 148 例が検出されている。正用法と認定される表現で 200 に届かない例は COCA において潜在的に多数存在している。本稿での主張は新たな正用法、あるいは非標準用法として英和辞典や語法書に記述する採用基準を考察したもので、既存の正用法が本稿の仮説を必ずしも満たす必要はない点を付け加えておきたい。

その他、ウィズダム英和辞典に記述のある非標準用法の **what do you got** は COCA では 1922 例検出され、正用法の **what have you got** (1509 例) を上回るような用例も見られる。本稿での仮説を満たす英語表現の記述が継続的に英和辞典・語法書に追記されていくことが今後も期待される。

6. まとめ

本稿では COCA を用いて非標準用法および正用法を新規認定する判断基準について検討を行った。対立する正用法が存在する逸脱表現については、基本的に英和辞典や語法書で記述されることはないが、本稿の仮説を満たす表現については、各仮説の判定に基づいて記述されることが望ましいと結論づける。ある一定数の使用が確認される言語現象については、英語学習者が辞典等で調べる可能性が十分あり、正用法だけでなく非標準の用法も例示することは英語学習上有益だと考えられる。本稿で調査したデータに基づくと、COCA は全体として会話表現の調査に対して有益な情報を提供してくれる傾向が伺える。COCA を使用する際は、調査対象の英語表現と対立する表現形式に注意を払いながら用法調査を行うことが肝要となる。

井上 (2010, 2019) は、辞書編纂におけるコーパス活用の一般的利点として次のものを指摘している。本稿での議論は (8a, f) と関連する。

(8) 辞書編纂におけるコーパス活用の一般的利点

- a. 先行研究にない言語事実についても分析記述が可能となる。
- b. 少人数のインフォーマント調査では避けることのできないインフォーマント自身の個人差を吸収できる。
- c. 母語話者の無意識の言語使用を分析の対象とすることが可能。
- d. 母語話者にはない視点での分析を可能にしたり、日本人英語学習者ならではの疑問点に答える文法・語法記述が可能となる。
- e. 一定の方針のもとに構築されたコーパスを使うことで、執筆者が水準を規定された共通の言語データに基づいて記述することが可能となる。
- f. コーパスを用いることで非文かどうかは判断できないが、一般的な表現であるかどうかはある程度判断ができる。

コーパスの中に複数の該当表現があったとしても、その表現が正用法かどうか断定できない点は 1 節の議論から明白であるが、当該表現が実社会において使用されているかどうかはコーパス調査からその実証を得ることができる。これまで記述されていない新たな英語表現の認定には、参照するコーパスの規模も影響するが、本稿で示した仮説を(8f)に関する具体的な判断基準として設定することで、新たな正用法の表現と非標準表現の記述が均質的に行えるようになる。

引用文献

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA)*. Available online at <https://www.english-corpora.org/coca/>.

Davies, M. (2015) Corpora: an introduction. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 11-31). Cambridge University Press.

井上永幸 (2010) 「辞書編集におけるコーパス活用」『英語語法文法研究』第 17 号: 5-22.

井上永幸 (2019) 「コーパスを活用した学習英和辞書編集の問題点」赤野一郎先生古希記念論文集編集委員会 (編)『言語分析のフロンティア』金星堂, pp. 48-60.

井上永幸・赤野一郎 (編) (2018)『ウィズダム英和辞典第 4 版』三省堂.

小西友七 (編) (1989)『英語基本形容詞・副詞辞典』研究社出版.

小西友七・南出康世 (編) (2001)『ジーニアス英和大辞典』大修館書店.

大橋浩 (2013) 「文法化」森雄一・高橋英光 (編)『認知言語学 基礎から最前線へ』くろしお出版, pp.155-179.

米国一般教書演説に出現する分離不定詞の効果に関する予備的研究

福本 広光(大阪大学 大学院生)

u163938a@ecs.osaka-u.ac.jp

A Preliminary Study on the Effect of Split Infinitives Appearing in the State of the Union Addresses

FUKUMOTO Hiromitsu (Osaka University, Graduate Student)

Abstract

The present study attempts to analyze the effect of several examples of split infinitives found in the transcripts of the U.S. State of the Union Address. We have examined the actual trend and the effectiveness of the construction when used in the political speech register. The examination suggests that the split infinitives sometimes function as a discourse strategy, emphasizing the Presidents' thoughts and claims.

Keywords

分離不定詞, 一般教書演説, 強意, ディスコース・ストラテジー

1. はじめに

アメリカ合衆国（以下米国）大統領が年頭に行う主要な演説の一つとして、一般教書演説（State of the Union Address）がある。米国行政府の長たる大統領が連邦議会に対して行う施政方針演説のことを指し、そこでは主に過去一年の現状や政治課題などが報告される。大統領の視点からは、政治要綱のすべてを一つの演説に盛り込んで伝える特別な機会であり、かつては憲法によって定められた義務から始まったものの、今や行政権や権威を行使するための原点にして、自らの権力について示す格好の場でもある。一方、連邦議会からも大統領の施政について詳細を知ることができる重要な演説の一つと目されている（Shogan (2015:1) の記述を基に）。

本研究は歴代米国大統領による一般教書演説の原稿をコーパスとして、そこに現れる分離不定詞（split infinitive）の用法の実態ないし効果について用例を解釈しながら探ることを試みる。分離不定詞は *to* 不定詞の間に副詞（相当語句）を嵌入させた構造（用例(1)参照）であり、その使用の可否については今なお議論を呼ぶ。規範的観点から、場面や状況によってその使用は制限されることもあるが、本研究を通して分離不定詞は一般教書演説という伝統ある演説の中でどのような振る舞いをしてきたのか、その一端に迫りたい。

- (1) We are working to finally end America's longest war and bring our troops back home.
(Trump, 2020)

2. 先行研究

分離不定詞に関する実証的な量的研究としては、多様な英米の通時的・共時的コーパスを活用した論考 (Fischer 2007; Mitrasca 2009; Calle-Martín and Miranda-García 2009; Perales-Escudero 2011; Calle-Martín 2015; 廣田 2021 ほか) に加え、近年ではいわゆる世界英語 (World Englishes) への関心が高まりつつあることもあり、アジア圏をはじめとした英語圏以外の地域で使用されている分離不定詞に関する研究も盛んである (Spakorn 2012; Calle-Martín and Romero-Barranco 2014; Gonzales and Dita 2018)。これらの研究の手法はおおよそ通ったものが多く、頻度の英米差 (などの地域差) やテキストジャンル間での頻度の比較, *to* 不定詞を「割る」要素である副詞に着目した分析, 同一コーパス上における分離不定詞とそれを避けた構文との頻度や割合の比較などを中心とした考察が行われている。

管見ではこれまで演説をデータソースとした分離不定詞研究はほとんど存在しなかったことから、拙論 (福本 2022) では米国一般教書演説をデータとした分離不定詞の記述的分析を行った。ここでは取り分け 1960 年代以降の大統領が用いた用例において分離不定詞と共起する表現に着目し、命令や依頼表現、列挙や並列項目の一つとなって接続詞 *and* と共起しやすいことなどを明らかにした。

3. リサーチデザイン

3.1 研究の目的

本研究は一般教書演説のデータを集積して作成したコーパスに基づき、“splitter” (*to* 不定詞を「割る」副詞相当語句, 以下 *splitter*) を軸として用例の文脈を分析し、その傾向や効果の実態の一部を炙り出す。そして演説レジスターにおいて分離不定詞を使用することによって得られる有用性についてアプローチを試み、今後の言語文化学的考察に繋げようとするものである。

3.2 データ・手法

調査においては、オンライン上のデータベースである *The American Presidency Project* 所収のスク립トテキストを用いる。一般教書演説の原稿も歴代のものが原則全て収録されており、一種の通史的コーパスとしての機能も果たすと考えられる。分析対象となるのは、George Washington によって 1790 年になされた最初の一般教書演説から Donald Trump によって 2020 年になされたものまで、44 名の大統領、約 230 年分 (1,825,708 語) の言語資料である。これらを一つずつテキストエディタに貼り付け、それらから“*to* + 副詞(相当句)+原形動詞”のフレーズを抽出すべく Brill’s Tagger にかけて一つずつ品詞標識を付与した。完成したデータセットを AntConc (4.1.1) にかけて “*to_TO *_RB *_VB**” “*to_TO *_JJ *_VB**” “*to_TO *_IN**” といった検索式で検索を行い、分離不定詞構造の用例とその前後に出現する文脈を抽出した。特に後者 2 つの検索式では、得られた用例が分離不定詞に該当するかについてコンコーダンスラインを目視し、手作業で分類を行った。

4. 結果と考察

4.1 一般教書演説における分離不定詞頻度の変遷

図 1 歴代大統領による一般教書演説中の分離不定詞頻度 (per 10,000 words)

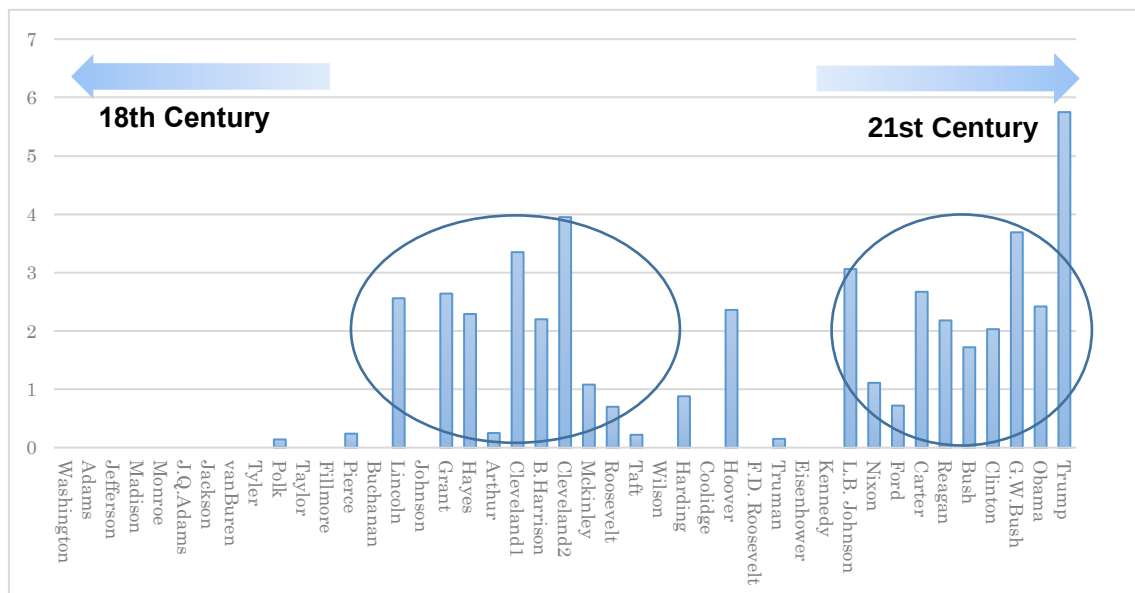


図 1 は先述の一般教書演説データセットに含まれる歴代大統領による分離不定詞の 1 万語当たりの調整頻度を表したものである。19 世紀半ばまで分離不定詞は演説内では避けられており、副詞が *to* 不定詞を修飾する場合は基本的に「文法的に適正な」位置に置かれていたが、Lincoln (1861 年就任) 以降、20 世紀初頭までに就任したほぼ全ての大統領は分離不定詞を演説内で用いている(図 1 中左円)。20 世紀前半にはその使用は再び下火になるも、後半期、即ち Lyndon Johnson 就任後の 1964 年以降は全ての人物が分離不定詞を任期内の演説に含めている (図 1 中右円)。このグラフからは分離不定詞使用に関しては時代ごと、あるいは人物ごとの頻度の揺れはあるものの、全体的に見ると一般教書演説においても一定数使用されていることがわかる。

4.2 一般教書演説における分離不定詞の効果について

本節では、一般教書演説において分離不定詞の実態とそれらを用いることで得られる効果について、図 1 にて円で囲った 2 つの時代に関してそれぞれ分析を行う。

4.2.1 米国創設期から 20 世紀初頭にかけての大統領による分離不定詞の効果

まず、Washington (1789 年就任) から Taft (1908-1912 年) までの期間に用いられた分離不定詞を扱う。ここで区切ったのは、最初に断続的に分離不定詞が確認できた期間のためである。実質的には 1860 年代以降の用例がほとんどを占めるが、それ以前に僅かながら用例が抽出されたため、それらもここに含めている。表 1 はこの期間に使用された *splitter* の一覧である。今回検出された全 211 例中 113 例が、この期間において使われている。

表 1 20 世紀初頭までの一般教書演説における分離不定詞 *splitter* (カッコ内はトークン数)

so (10x); *further* (8x); *fully, seriously* (5x); *properly* (4x); *again, at once, carefully, earnestly, especially, personally* (3x); *adequately, annually, better, entirely, judiciously, longer, soberly, such as* (2x); *about, above, at least, best, broadly, constantly better, dangerously, decently, disastrously, duly, easily, effectively, effectually, even, exactly, expressly, finally, freely, greatly, habitually, here, honestly, intelligently, more surely, more than, not, now, otherwise beneficially, partially, particularly, peaceably, persistently, promptly and effectively, ruinously, silently, simultaneously, somewhat, still more materially, still, strictly, strongly, thoroughly, thus, unduly, vastly, very justly, wrongfully* (1x)

副詞群を観察すると, *further, fully* をはじめとした「程度」の意味を表す副詞が頻繁に用いられているほか, *earnestly* (切に), *very justly* などの「強意」の意味を表す副詞や「様態」の副詞も一部含まれている。 *decently, judiciously, peaceably* など一般的には普遍性が低いと考えられる語や *so, now, here, about* など, 普遍性は高いが現代の分離不定詞においてはほとんど使用されないような副詞も *splitter* として使用されていることが特徴的である。

確認された分離不定詞の中には, 例えば *It is of the first importance* と共起して国家の中で永続する部分(領土)を「正しく」評価することが何より大切なことだ, とする(2)のようなものや, (3)のように *I desire* と共起して国債発行に関する現行の法令を改正することの賢明さを「強く」勧めたい, という文脈の, 大統領の意見・主張を強調する場面で用いられていると判断される用例も一定数みられたが, 一方で(4)や(5)のような単に事実を誤解なく伝えるという目的で利用されていると判断せざるを得ないものも多くあった。

- (2) **It is of the first importance to duly consider and estimate** this ever-enduring part.
(Lincoln, 1862)
- (3) **I desire** also **to earnestly suggest** the wisdom of amending the existing statutes in regard to the issuance of Government bonds.
(Cleveland2, 1893)
- (4) The examinations have been attended by many citizens, who have had an opportunity **to thoroughly investigate** the scope and character of the tests and the method of determining the results, and those visitors have without exception approved the methods employed, and several of them have publicly attested their favorable opinion.
(Hayes, 1879)
- (5) It is imperative that much of the money provided for such construction should now be allotted to the erection of necessary quarters for the garrisons assigned to the coast defenses, where many men will be needed **to properly care for and operate** modern guns.
(Cleveland2, 1896)

この時代の一般教書演説は文章が書き言葉的であり、かつ一文が長い用例が多く抽出された。その場合には分離不定詞を避けるよりも使用した方が却って文意が明快になる、といった機能的側面から使用されていた可能性もある。いずれにせよこの時代における用例のより精緻な分析が喫緊の課題である。

4.2.2 20 世紀後半以降の大統領による分離不定詞の効果

続いて、20 世紀後半、すなわち 1960 年以降、2020 年に至るまでの一般教書演説において用いられている分離不定詞の効果や用法について再考する。今回検出された全 211 例中、90 例がこの期間において使われている（福本（2022: 62）の調査では複数語 *splitter* および Reagan 1981 のデータを含めていないため、一部数値が本稿と異なっている場合がある）。表 2 は調査にて得られた *splitter* の一覧である。

表 2 20 世紀後半期以降の一般教書演説における分離不定詞 *splitter* (カッコ内はトークン数)

<i>finally, fully</i> (8x); <i>better, further</i> (6x); <i>just, ultimately</i> (4x); <i>always, quickly, successfully</i> (3x); <i>carefully, greatly, immediately, more than, not only, permanently, really</i> (2x); <i>actually, again, aggressively, all, also, annually, at least, best, clearly, closely, constantly, dramatically, effectively, even, ever, faithfully, fully and suddenly, instantly, massively, never, not, only, promptly, safely, specifically, substantially, swiftly, totally, truly, voluntarily</i> (1x)

4.2.1 節で見た副詞同様、繰り返し用いられているのは *fully, further, better* など「程度」を表す副詞が中心であり、「様態」の副詞は *immediately* や *promptly* などの「速さ」を表すものを除いては多くない。全体を概観すると *just, really, greatly* など「強意」の意味を持つ副詞も頻繁に用いられ、その多くはいかなるジャンルの英文においても普遍的に用いられる副詞で構成されている。

この時代には、明らかに大統領の主張や思惑を強調することを目的として分離不定詞が用いられていると判断される用例が収集された。

- (6) **I am therefore recommending** that the one-year investment tax credit of 12 percent be extended an additional two years **to specifically speed** the construction of power plants that do not use natural gas or oil. (Ford, 1975)
- (7) I want to push education reform, as I said, **not just to** spend more money **but to really improve** learning. (Clinton, 1993)
- (8) And **the time has come** to put the national interest above the special interest and **to totally eliminate** political action committees. (Bush, 1991)
- (9) Every Federal dollar **should be** leveraged by partnering with State and local governments and, where appropriate, tapping into private sector investment -- **to permanently fix** the infrastructure deficit. (Trump, 2018)

(6) では天然ガスや石油を使わない発電所の建設を早めるという大統領の思惑を強調している。(7)では *not just...but* 構文と共起し、さらに *really* を伴わせることによって、教育改革を進めたいのは「学び」を改善するため、という主張点を強化する役割を果たしている。(8) や (9) においては *totally* や *permanently* と、強意語の中でもかなり意味合いが強く、大袈裟にも思われるような副詞を埋め込んだ分離不定詞を使用することにより、(8)では政治行動委員会の廃止の必要を強く主張することができ、(9)では連邦ドルを地方政府との提携によって、民間部門の投資も活用しつつ活用すべきであるという大統領の主張はインフラの赤字の解決が最終目的であることを聴衆に強く印象付ける。特に(8)では、*the time has come* という表現と共起し、あたかも聴衆を扇動するような文脈が成り立っている。この時代に演説を行った大統領は、科学技術の発達による伝達手段の増加と共に、世界中の様々な階層の聴衆に適切に演説内容を伝える必要に迫られた。その際に有標な構造である分離不定詞を敢えて用いることで、自身の主張点を明らかにするためのディスコース・ストラテジーの一つとして機能させたことが示唆される。

5. まとめ

本研究では、米国政治において主要な演説である一般教書演説に出現する分離不定詞使用に伴う効果について予備的な検討を行い、未だ分析には改良の余地を残すものの、当構造が大統領の意志や主張を強化するためのディスコース・ストラテジーを形成しつつあることを示唆した。

今後、共起表現の観察や *splitter* の特質のみに拠らない分析の精緻化（特に歴史言語学や語用論の観点などから）に努めるほか、背景となる出来事や文化についての知識と絡めることによってコンテキストの「読み」を深化させ、分離不定詞使用の動機に関する考察を深めてゆきたい。

引用文献

- Calle-Martín, J. (2015). "The Split Infinitive in Middle English." *North-Western European Language* 68 (2): 227-259.
- Calle-Martín, J. and Antonio, Miranda-García. (2009). "On the Use of Split Infinitives in English." In Antoinette, R and Andrew, K (eds.) *Corpus Linguistics: Refinements and Reassessments*. 347-364. Amsterdam: Rodopi.
- Calle-Martín and Romero-Barranco. (2014). "On the Use of the Split Infinitive in the Asian Varieties of English." *Nordic Journal of English Studies* 13 (1): 130-147.
- Fischer, R. (2007) "To Boldly Split the Infinitive or Not? Prescriptive Traditions and Current English Usage." In Stephan E., Nils L, Joachim S. and Wim V.(eds.) *Germanic Language Histories "from Below" (1700-2000)*, 259-273. Berlin: Walter de Gruyter.
- 福本広光 (2022). 「米国一般教書演説における分離不定詞についての記述分析：1960年代から2020年の用例を中心に」『言語文化共同研究プロジェクト2021』57-68. 大阪大学.
- Gonzales, Wilkinson Daniel Wong and Shirley N. Dita. (2018). "Split Infinitives across World Englishes: A Corpus-based Investigation." *Asian Englishes* 20 (3): 242-267.
- 廣田友晴 (2021). 「否定分離不定詞の変遷」『語法と理論との接続をめざして』187-208. ひつじ書房.
- Mitrasca, M. (2009). "The split infinitive in electronic corpora: Should there be a rule?" *Concordia Working Papers in Applied Linguistics*, 2: 99-131.
- Perales-Escudero, M. D. (2011). "To Split or to Not Split: The Split Infinitive Past and Present." *Journal of English Linguistics* 39 (4): 313-334.
- Shogan, C.J. (2015). "The President's State of the Union Address: Tradition, Function, and Policy Implications." (<https://fas.org/sgp/crs/misc/R40132.pdf>)
- Supakorn, P. (2012). "The English split infinitive: A comparative study of learner corpora" *International Journal of Research Studies in Language Learning*. 2 (4): 21-32.

憲法をめぐる日本の帝国議会・国会会議録における「翻訳」の用法および 米国議会会議録における translation の用法の分析

島津 美和子(立教大学アメリカ研究所)

miwako-rikkyo@rikkyo.ac.jp

Analysis of "*honyaku*" used in the Japanese Imperial and National Diet minutes and "translation" used in the U.S. Congressional Record in the context of constitutional law

SHIMAZU Miwako (Institute for American Studies,
Rikkyo University)

Abstract

This study analyzes the use of *honyaku* [翻訳 translation] in the minutes of the Japanese Imperial and National Diets in the context of constitutional law. We examine the validity of the contention that Japan's current national constitution is a translation of an original English draft by the General Headquarters of SCAP and was not drafted by the Japanese. We conclude that Japanese National Diet members more often associate the Constitution with *honyaku* and use the term with the suffix *cho* or *shu*, which imparts a negative connotation, than did Imperial Diet members. As a contrastive study, we further analyze the use of "translation" in a national constitutional discourse in the U.S. Congressional Record. Surprisingly Congress members, while expressing general appreciation of translation of the constitution, make no direct reference to the Japanese case.

Keywords

翻訳, translation, 憲法, 翻訳調

1. はじめに

現在、日本政府が進める改憲の理由の一つに、現行憲法は GHQ から提示された英語原案を日本語に翻訳したものであり、日本が策定したものではないという主張がある。例えば、自民党が 2015 年に発表した政策パンフレット『ほのぼの一家の憲法改正ってなあに？』では、孫の「憲法って変な日本語が多くない？」という疑問に対し、祖父が「今の憲法の前文にはアメリカ合衆国憲法やリンカーンの演説なんかが翻訳口調で使われているからなあ」「日本国憲法の^{もと}基を作ったのがアメリカ人だからじゃよ」と答える場面がある。祖父の主張は、換言すると、現行憲法の前文はアメリカ人が作成したもので、それを翻訳した現行憲法は翻訳の口調が残り、質が悪いということだ。ここ

では翻訳が否定的な意味で用いられている。この主張の言語的妥当性を検証するため、本研究では、帝国議会・国会会議録をコーパスとみなし、憲法の文脈における「翻訳」の用例を抽出し、分析する。その比較対象として GHQ の主体となった国の議会として、米国議会会議録内の憲法の文脈における translation の用例を抽出し、同様に分析する。これらの分析結果をもとに先の主張の妥当性について考察を行い、結論を提示する。

2. 先行研究

研究テーマと研究データの 2 つの観点から述べる。

第一に、「はじめに」で言及した翻訳の口調は、形態的には「翻訳調」として用いられることが多い。「翻訳調」を主要テーマとする論考は、タイトルまたはキーワードに「翻訳調」を含むものに絞ると、日本語文献には数少なく、国文学と通訳翻訳研究で各 1 件ある(伊原, 2000; 梅林, 2000)。しかし、これらは文学を対象とし、法令文や技術文書を対象としていない。一方、英語文献では、「翻訳調」の英語相当表現 translationese をタイトルまたは要約に含む論考は一定数存在する。そのうち日本語について論じた論考には文学を対象とした 3 件(Meldrum, 2009a, 2009b, 2009c)と翻訳調の受容度に関する調査報告 1 件(Furuno, 2005)がある。同調査報告は翻訳者を対象に翻訳調の逐語的な日本語と自然な日本語のどちらを好むかを調査したもので、後者の率が高いことを報告している。

ここで辞書上の語義を確認しておく。「翻訳調」を立項しているのは『広辞苑』と『大辞林』で、それぞれ「外国語を日本語に訳してできたような、それまで日本語では使わなかった表現や文体」(2018, p. 2733), 「外国語を日本語に直訳したような独特の表現。また、そのような文体の作品」(2019, 下, p. 2407)とする。これより、辞書編纂者は「翻訳調」の語自体には否定的な意味を付与していないことが分かる。ただし、これらの語釈は、具体的にどのような表現をもって外国語を日本語に翻訳した文章の特徴とみなすのかについては答えていない。一方, translationese の OED の語義は「(悪い)翻訳の性質と感じられる言語様式。翻訳文書の中のぎごちない, 不自然, または, 慣用的でないようにみえることば, 特に翻訳者が原文に特有の特徴を訳文にできるだけ持ち込むように試みた結果, 生じたことば」(以上, 筆者訳)となっている。否定的な意味が認められる。

第二に、国会会議録を「戦後 60 年近くにわたる...話し言葉での膨大な発話記録」(松田, 2008, p. 1)あるいは大規模な日本語データ(松田, 2012, p. 55)と捉え、言語研究に生かすことがなされてきた。国会会議録検索システムが 2014 年から Web API として提供されるようになり、利便性が増したことも関係していよう。しかしながら、帝国議会会議録・国会会議録の双方を言語研究に使用した例は管見の限りこれまでない。

3. リサーチデザイン

3.1 研究目的と研究設問

本研究は、現行憲法は GHQ から提示された英語原案を日本語に翻訳したものであり、日本が策定したものではないという主張の言語的妥当性を検証することを目的とし、研究設

問を「日本の憲法の文脈において翻訳が国会の場でどのように語られてきたか。また、日本国憲法成立後と成立前で語られ方に違いがあるか」と設定する。

3.2 データ

3.2.1 日本語データ

帝国議会議録および国会議事録から成る。前者については、帝国議会議録検索システム(<https://teikokugikai-i.ndl.go.jp>)に対し、キーワードに「帝国憲法改正 AND 翻訳」を設定し検索し得られた 36 会議の会議録のうち、憲法の文脈で「翻訳」を用いた発話を日本語データに使用する(表 1)。表 1 に示す会議録は帝国議会における帝国憲法改正案の審議の過程を記録したものであり、これらの会議録内の発話は帝国議会改正案をめぐる議員が「翻訳」をどう語っているかを見るために使用する。衆貴両院の帝国憲法改正案委員会と帝国憲法改正案委員小委員会の会議録が大半を占める。これらの会議では、各議員から出された意見

表1: 使用した帝国議会議録

帝国 議会 回次	院名	会議名	号数	開会日付
1	90 衆	本会議	5号	昭和21年6月25日
2	90 衆	本会議	6号	昭和21年6月26日
3	90 衆	本会議	8号	昭和21年6月28日
4	90 衆	帝国憲法改正案委員会	2号	昭和21年7月1日
5	90 衆	帝国憲法改正案委員会	4号	昭和21年7月3日
6	90 衆	帝国憲法改正案委員会	6号	昭和21年7月5日
7	90 衆	帝国憲法改正案委員会	7号	昭和21年7月6日
8	90 衆	帝国憲法改正案委員会	10号	昭和21年7月11日
9	90 衆	帝国憲法改正案委員会	11号	昭和21年7月12日
10	90 衆	帝国憲法改正案委員会	13号	昭和21年7月15日
11	90 衆	帝国憲法改正案委員会	15号	昭和21年7月17日
12	90 衆	帝国憲法改正案委員会	17号	昭和21年7月19日
13	90 衆	帝国憲法改正案委員会	18号	昭和21年7月20日
14	90 衆	帝国憲法改正案委員小委員会	1号	昭和21年7月25日
15	90 衆	帝国憲法改正案委員小委員会	2号	昭和21年7月26日
16	90 衆	帝国憲法改正案委員小委員会	3号	昭和21年7月27日
17	90 衆	帝国憲法改正案委員小委員会	4号	昭和21年7月29日
18	90 衆	帝国憲法改正案委員小委員会	7号	昭和21年8月1日
19	90 衆	帝国憲法改正案委員小委員会	8号	昭和21年8月2日
20	90 衆	帝国憲法改正案委員小委員会	12号	昭和21年8月16日
21	90 貴	本会議	23号	昭和21年8月26日
22	90 貴	本会議	24号	昭和21年8月27日
23	90 貴	帝国憲法改正案特別委員会	1号	昭和21年8月31日
24	90 貴	帝国憲法改正案特別委員会	4号	昭和21年9月4日
25	90 貴	帝国憲法改正案特別委員会	5号	昭和21年9月5日
26	90 貴	帝国憲法改正案特別委員会	6号	昭和21年9月6日
27	90 貴	帝国憲法改正案特別委員会	8号	昭和21年9月9日
28	90 貴	帝国憲法改正案特別委員会	9号	昭和21年9月10日
29	90 貴	帝国憲法改正案特別委員会	12号	昭和21年9月13日
30	90 貴	帝国憲法改正案特別委員会	14号	昭和21年9月16日
31	90 貴	帝国憲法改正案特別委員会	15号	昭和21年9月17日
32	90 貴	帝国憲法改正案特別委員会	17号	昭和21年9月19日
33	90 貴	帝国憲法改正案特別委員会	22号	昭和21年9月26日
34	90 貴	帝国憲法改正案特別委員小委員会筆記要旨	3号	昭和21年10月1日
35	90 貴	帝国憲法改正案特別委員会	24号	昭和21年10月3日

表2: 使用した国会議事録

国会 回次	院名	会議名	号数	開会日付
1	22 衆	内閣委員会	49号	昭和30年7月28日
2	24 衆	内閣委員会	22号	昭和31年3月13日
3	24 参	内閣委員会	40号	昭和31年5月9日
4	68 衆	内閣委員会	15号	昭和47年4月25日
5	71 衆	法務委員会	46号	昭和48年9月21日
6	87 参	内閣委員会	13号	昭和54年5月31日
7	126 衆	政治改革に関する調査特別委員会	3号	平成5年3月17日
8	143 衆	予算委員会	4号	平成10年8月19日
9	147 衆	憲法調査会	4号	平成12年3月9日
10	147 衆	憲法調査会	8号	平成12年4月27日
11	149 衆	憲法調査会	1号	平成12年8月3日
12	150 衆	憲法調査会	3号	平成12年10月26日
13	150 衆	文教委員会	4号	平成12年11月17日
14	151 衆	憲法調査会	7号	平成13年6月14日
15	156 参	憲法調査会	7号	平成15年5月14日
16	160 衆	憲法調査会	1号	平成16年8月5日
17	161 参	本会議	2号	平成16年10月14日
18	161 衆	憲法調査会公聴会	1号	平成16年11月11日
19	161 衆	憲法調査会公聴会	2号	平成16年11月18日
20	161 参	憲法調査会	6号	平成16年12月1日
21	161 衆	憲法調査会	4号	平成16年12月2日
22	162 衆	憲法調査会	3号	平成17年2月17日
23	162 衆	憲法調査会	4号	平成17年2月24日
24	179 参	憲法審査会	3号	平成23年12月7日
25	180 衆	憲法審査会	1号	平成24年2月23日
26	180 参	憲法審査会	6号	平成24年5月30日
27	183 衆	憲法審査会	9号	平成25年5月16日
28	183 参	憲法審査会	5号	平成25年6月5日
29	183 衆	憲法審査会	12号	平成25年6月13日
30	187 参	憲法審査会	2号	平成26年10月22日
31	187 衆	憲法審査会	3号	平成26年11月19日
32	189 参	予算委員会	20号	平成27年8月24日

をもとに枢密院にて可決された憲法改正草案に対し各種の追加や修正がなされた。

一方、後者については、国会会議録検索システム(<https://kokkai.ndl.go.jp>)に対し、キーワードに「(翻訳調 OR 翻訳臭) AND 憲法」を設定し検索した結果、得られた会議のうち、憲法の文脈で「翻訳調」または「翻訳臭」を用いた発話を日本語データとして使用する(表 2)。文字列としては「翻訳調」と一致するが「翻訳調査」など「翻訳調」を意味しないものや刑事訴訟法など憲法以外の法令を指して「翻訳調」「翻訳臭」と述べているものは除外してある。表 2 内の発話は現行の日本国憲法と翻訳の関係について議員がどう語っているかを見るために使用する。検索対象を「翻訳」ではなく「翻訳調」「翻訳臭」に限定したのは、帝国議会議録でまとまった数、生起していた「翻訳調」,「翻訳臭」に限定することによって、目視で分析可能な数に抑えるためである。

3.2.2 英語データ

アメリカ議会図書館の運営する Congress.gov で公開されている 56 議会(1899-1901)~103 議会(1993-94)および 104 議会(1995-)以降の Congressional Record に対し, "Constitution translation"および"translation of * Constitution"を検索語とし, 当該サイトの検索システムで検索し得られた結果を 20 会議分の英語データとして使用した。検索時, 検索語はその派生語も含めて検索するよう条件設定した。なお, translationese の用例は文脈に関わらず皆無であった。

3.3 手法

3.3.1 データの事前処理

3.2 のデータは発話単位で Excel に集計し, 発話者名, 発話者出自, 該当会議名, 日付などの項目を追加した上で, 分析の効率化のためにテーブルに変換した。発話数は, 帝国議会 76, 国会 35, 米国議会 22 であった。

3.3.2 分析の手順

各発言の前後を精読し, 発話者がどのような背景と意図で「翻訳」(帝国議会の場合), 「翻訳調」あるいは「翻訳臭」(国会の場合), translation(米国議会の場合)を用いているかを吟味する。その際, 前後に出現する特徴的な語の有無も精査する。

4. 結果と考察

4.1 帝国議会議録

帝国議会議録中の「翻訳」を使用目的別に分類した結果を表 3 に示す。翻訳の行わないし翻訳結果を指し示す場合が約 8 割を占め, この場合, 発話者は翻訳に対して中立的であった。批判に用いる場合は, 「翻訳調」よりもむしろ「翻訳(の)口調」と表現することが多かった。また, 特に厳しく批判する場合は, 「極めて」「甚だ」「全く」といった程度副詞と共に起していた。

表3: 帝国議会会議録中の「翻訳」

		実際の表現	程度副詞 が前接
翻訳した結果の文書を 指し示すため	51	翻訳(42)、翻訳語(5)、 翻訳文(3)、翻訳物(1)	-
行為としての翻訳を指 し示すため	32	翻訳(32)	-
発話者本人が憲法前文 または憲法全体を批判 するため	19	翻訳(の)口調(11)、翻訳 的(3)、翻訳臭(2)、翻訳 の印象(1)、翻訳文の感 (1)、翻訳の臭い(1)	3
他者の憲法前文または 憲法全体に対する批判 に言及するため	2	翻訳的(1)、翻訳調(1)	1
翻訳主体	1	翻訳者(1)	-
延べ数	105	-	4

表4: 国会会議録中の「翻訳調」「翻訳臭」

		程度副詞 が前接	強意語が 前接	翻訳臭さ
発話者本人が憲法前文 または憲法全体を批判 するため	21	5	0	1
他者の憲法前文または 憲法全体に対する批判 に言及するため	9	1	0	0
発話者本人が自党を代 表して憲法前文または 憲法全体を批判するた め	6	2	1	0
延べ数	36	8	1	1

4.2 国会会議録

国会会議録中の「翻訳調」「翻訳臭」を使用目的別に分類した結果を表 4 に示す。発話者が自ら憲法を批判するケースが過半数を占め、発話者が自党を代表して批判するケースと合わせると 7 割強になる。このうち 2 割強は、極めて、非常に、余りに(も)、「切りがないくらい」などの程度副詞を用い、手厳しく批判していた。また、批判する発話者は「翻訳調」を補完する目的で、しばしば言い換え表現を使っていた。わかりにくい・わかりづらいが最も多く、このほか「悪文、意味がはっきりしない、英語みたい、簡潔性・明確性に欠ける、国語としてなっていない、生硬、日本語としてきわめつきの悪文、日本語としてまずい、日本語になじまない語感をもつ、日本語らしくない、美しくない、不正確、模範的な日本語ではない、わかりにくい表現を多用」があった。

4.3 米国議会会議録

表5: 米国議会会議録中の Constitution と共起する translation

合衆国憲法の他の言語 への翻訳	10	子ども向け(5)、スペイン語 訳(3)、ハンガリー語訳 (1)、ギリシャ語訳(1)
他国の憲法の英訳	7	キューバ憲法(3)、ニカラ グア憲法(2)、ブラジル憲 法(1)、ロシア憲法(1)
他の言語への翻訳以外 の意味	5	
延べ数	22	-

米国議会会議録中の translation(派生語を含む)を内容別に分類した結果を表 5 に示す。合衆国憲法の他の言語への翻訳を指す場合が最も多く、次いで、他国の憲法の英訳を指す場合が多くみられた。前者の場合、どのケースも合衆国憲法の他言語訳およびその翻訳者を高く評価している点が特徴的であった。一方、後者については、該当国の憲法の英訳は、米国との貿易関係

や外交関係上、その国の国家基盤を把握するために使われていた。議員が該当国の言語を理解できないためである。この場合、議員は翻訳に対して否定的でも肯定的でもない。このほかに、日本語では「翻訳」に相当しない translation の用例もあった。これらは日本語の「変える」「移す」などが当てはまる。例えば、"translate [the Constitution] into real life"がある。この場合、いずれも Constitution は合衆国憲法を指し、into が後続する点が共通していた。一方、日本国憲法は今回のように翻訳の文脈に限定すると用例が皆無であった。さらに、日本の国会で盛んに取り上げられた日本国憲法の英語版についても言及がみられなかった。

5. まとめ

本論考では、現行憲法は GHQ の英語原案の日本語訳であり、日本が策定したものではないという主張の言語的妥当性を検証するために、日本の憲法の文脈において翻訳が国会の場でどのように語られ、また、日本国憲法成立後と成立前でその語られ方に違いがあるかという研究設問を

立て、議員の発言内容を分析した。帝国議会および国会での議員の発言から「翻訳」を抽出し、分析した結果、帝国議会においては憲法の日本語原案、国会では現行憲法に対して翻訳の調子を持つとして否定的または批判的な見解がみられた。加えて後者ではその否定の強度が増し、また議員が自党を代表して批判するケースが新たに生じた。対照事例として、米国議会において翻訳が憲法の文脈でどのように語られてきたかを同様に分析した結果、合衆国憲法の他言語訳に対しては肯定的であるが、それ以外は中立的であることをみた。言語や国の立場の違いも関与するが、日本の状況は必ずしも他国ではみられないことがわかった。一方で、広く日本の論壇に目を向けると法令文や技術文書の翻訳について「翻訳調」として、しばしば問題になる「日本語」は、実は「科学」「社会」など無数の、外来の諸概念を実現するための新造語の創出とともに、それを可能にするための日本人の生み出した貴重な工夫であったと考えるべきであろう(村上, 2021, p. 96)といった前向きな見方もある。日本国憲法は翻訳調という紋切型の議論は再考の必要があると思われる。

引用文献

- Furuno, Y. (2005). Translationese in Japan. In E. Hung (Ed.), *Translation and cultural change: Studies in history, norms and image-projection* (pp. 147-160). Amsterdam: John Benjamins.
- 伊原紀子(2000)「文学翻訳における異化・同化」『国際文化学』(神戸大学国際文化学会), 03, 105-116.
- 松田謙次郎(2008)「国会会議録検索システム概論」松田謙次郎編『国会会議録を使った日本語研究』(pp. 1-32) ひつじ書房.
- 松田謙次郎(2012)「国会会議録をつかう」日比谷潤子編『はじめて学ぶ社会言語学—ことばのバリエーションを考える 14 章』(pp. 54-79) ミネルヴァ書房.
- 松村明 三省堂編修所編(2019)『大辞林』第 4 版 三省堂.
- Meldrum, Y. F. (2009a). *Contemporary translationese in Japanese popular literature* [Doctoral dissertation, University of Alberta].
- Meldrum, Y. F. (2009b). Translationese in Japanese literary translation. *TTR*, 22(1), 93–118. DOI: doi.org/10.7202/044783ar
- Meldrum, Y. F. (2009c). Translationese-specific linguistic characteristics: A corpus-based study of contemporary Japanese translationese. 『通訳翻訳への招待』3 号, 111-131.
- 村上陽一郎 (2021) 『文化としての科学/技術』 岩波書店.
- 新村出編(2018)『広辞苑』第 7 版 岩波書店.
- "translationese, n." *OED Online*. Oxford University Press, June 2022. Web. 15 August 2022.
- 梅林博人(2000)「近代小説にみる接続詞「そして」 翻訳調といわれる「A そして B」をめぐる」『国文学 解釈と鑑賞』, 65-07, 36-42.

日本人英語学習者の「うなぎ文」の使用に関する分析

藤原 康弘(名城大学)
岩男 考哲(神戸市外国語大学)
伊藤 創(関西国際大学)
仲 潔(岐阜大学)
fujiwara@meijo-u.ac.jp

An Analysis on the Use of *Unagi* -sentences by Japanese Learners of English

FUJIWARA Yasuhiro (Meijo University)
IWAO Takanori (Kobe City University of Foreign Studies)
ITO Hajime (Kansai University of International Studies)
NAKA Kiyoshi (Gifu University)

Abstract

This study analyzed 1) the use of so-called *unagi*-sentences in the English written and spoken by Japanese learners of English, and 2) changes in the use of these sentences with their English proficiency increased. The results showed that the frequency of *unagi*-sentences, whether spoken or written, may not be high and will not change significantly with their second language proficiency. These results were examined not only from the perspective of second language acquisition and English language teaching, but also from the perspective of World Englishes.

Keywords

unagi-sentences, topic-comment constructions, L1 transfers,

1. はじめに

言語には、それぞれ特有の事態描写の型が存在する。英語と日本語でいえば、英語は「名詞中心構文」であるのに対して日本語は「動詞中心構文」(外山,1973, 外山,1992), 英語は「人間中心」で日本語は「状況中心」(国広,1974; Hinds,1986), 英語は「スル的表現」を好み日本語は「ナル的表現」を好む(国広,1974; 池上,1981), 英語は「ダイアログ言語」で日本語は「モノログ言語」である等々、さまざまな違いが指摘されている。

こうした事態描写の型の違いが事態把握の型の違いにも影響を与えることも、いわゆる言語相対論の検証という形で多く示されている(Slobin,1987; Boroditsky,2001 等)。したがって母語(以降 L1)は、学習言語(以降 L2)使用の際の認知的な事態把握や描写の違いにも表れうる(伊藤,

2018; 今井, 2020), また, その使用者のアイデンティティにも大きく影響する(Hino, 2021)。

このように考えれば, L2 に L1 の影響が見られること, 例えば日本語を母語とする英語学習者 (Japanese learners of English, 以降 JLE) による英語使用に日本語の事態把握・描写の影響がみられることはごく自然なことであろう。長い間, 単一言語話者の規範や「理想の母語話者」に固執してきた歴史を持つ第二言語習得研究も, 多言語社会への認識が高まるにつれ, 徐々に multilingualism に理解を示し, 近年では multilingual turn が起こっている (May, 2013)。

このような流れにあって, (他言語はともかく少なくとも) 英語学習においては, 母語の影響は肯定的に捉えられつつあるが, その良し悪しの境界線について明確な結論はないのが現状である。また言語ごとの事態描写¹の型の違いについての記述は数多くあるものの, 1) 異なる型のどの部分の習得が容易で, どの部分が困難か, 2) それはなぜか, そして 3) 異なる事態把握, 把握の型がどのような過程を経て習得されていくのか, については不明な点が多い。

こうした状況に鑑み, 筆者らはこれらの疑問を明らかにするために, JLE の英語使用, およびさまざまな言語を母語とする日本語学習者の日本語使用について, 学習者コーパスを用いて, その型の習得のあり方について特に動詞の自他に焦点をあて分析を行った。一般に日本語では自動詞が, 英語では他動詞が好まれるとされるが (木村, 1993, Hinds, 1986 等多数), 筆者らの調査によると, JLE の英語使用において自他の使用比率は英語母語話者とほぼ変わらず, 学習者のレベルが A2 レベルから B2 レベルに達しても, 特に変化は見られなかった (伊藤・岩男・藤原・仲, 2022a)。また, 日本語母語話者は, その事態を引き起こした Agent を背景化して描くような自動詞表現 (ドアが「開く」, 授業が「始まる」, 話が「まとまる」, など) を数多く用いるが, JLE の英語使用にこれらの表現はほとんど見られず, 学習者のレベルによってもその使用比率に差は見られなかった (伊藤他, 2022a)。さらに興味深いことに, 日本語学習者の日本語使用においても同様の結果が得られた (伊藤・藤原・岩男・仲, 2022b)。

これらの結果から, 1) L1 において多用される表現が必ずしも L2 で多用されるわけではなく, 2) L1 の表現の型に即した表現であっても L2 の対応表現の習得が容易とは限らないこと, そして, 3) L1 が何であるかにかかわらず習得の困難な項目があるということ, が示される。

2. 先行研究

2.1 うなぎ文とは

上述の先行研究の背景を踏まえ, 本研究では, 新たに名詞／形容詞述語文に焦点をあて, L1 の事態把握の型が L2 においてどのように現れるかについて調査を行う。

本発表で焦点を当てるのは「うなぎ文」²と呼ばれる文である。日本語は英語と異なり「名詞＋だ (です)」という形式で述語となることが可能である。これを名詞述語文と呼ぶ。具体例として以下が挙げられる (下線部が述語)。

- (1) 岩男君は甘党だ。
- (2) リーダーは仲君だ。
- (3) 藤原君は『海辺のカフカ』だ。

ここでは(1) (2) の詳細はおくとして(これについては伊藤・藤原・岩男・仲, 2022b を参照), (3) に注目したい。これが「うなぎ文」と呼ばれる表現の例である。(3) で「は」が提示する名詞「藤原君」と述語である「海辺のカフカ(だ)」との間には(1) (2) と違って, 直接的な論理関係が存在しないことが分かる。これが「うなぎ文」の主な特徴である。

この文をめぐるのは従来, その成立条件に議論が集中しがちであったが, うなぎ文の指す範囲についての言及も僅かながら存在する(益岡・田窪, 1992)。そこではうなぎ文は名詞述語文に留まらないという指摘がされている。例えば以下のようなものがそれにあたる。

(4) 伊藤君はぶかぶかだ。

本発表では, 今後その考察対象をより広げていくことを念頭に置いて, (4)のように名詞以外を述語にとるケースもうなぎ文として考察する。具体的には名詞述語文だけでなく, 形容詞述語文も分析対象とする。

2.2 研究課題

この日本語のうなぎ文の特徴は, JLE が使用する英語に転移すると報告されてきた(e.g., Sasaki, 1990; Shibata, 2006)。たとえば “Today is busy.”や “I’m tea.”などである。その主語と述語の意味関係の透明性によるものの, 多言語主義や国際英語論の観点から, 日本語母語話者が使用する英語のスタイルの一つとして積極的に認めてもよいという主張もある(Honna, 2010; 橘, 2014; 小宮, 2016; D’Angelo, Yamaguchi & Fujiwara, 2021)。しかしながら, 実際に日本語母語話者が使用する英語にどの程度うなぎ文の使用が見られるのか, 特に能力の伸長による変化の詳細は明らかではない。そこで本研究では, 1) JLE の英語におけるうなぎ文の使用, および 2) 能力の伸長による使用傾向の変化について分析する。

3. 研究手法

3.1 データ

分析で用いたコーパスは The International Corpus Network of Asian Learners of English (以降, ICNALE, Ishikawa, 2013) の日本人英語学習者の書き言葉エッセイ(Written Essays, ver. 2.4), および話し言葉モノログ(Spoken Monologues, ver. 2.0)のデータである。ICNALE は神戸大学の石川慎一郎氏が牽引して構築された英語学習者コーパスで, 日本, 韓国, 台湾, 中国, フィリピン, マレーシアなどの 10 のアジア諸国／地域出身の英語学習者のデータで構成されている。トピックを (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 つのみに統制することで, 学習者の能力(CEFR A2, B1, B2)や母語／出身地(日本, 中国, 韓国, パキスタンなど)の要因を用いた精緻な比較を可能にしている。本分析では JLE の書き言葉エッセイと話し言葉モノログの A2, B1, B2 以上の 3 つの能力層のデータを用いた。

3.2 分析手法

分析手法は、まず *AntConc* (ver. 3.5.9) を用いて、日本人英語学習者のデータから日本語の名詞述語文、形容詞述語文に該当する次のパターンを抽出した (BE+DT/NN*/JJ*)。すなわち、am/are/is/was/were/be/being/been + 決定詞 (a, an, the, no, this, that…) ～／名詞 (単数・複数)／形容詞 (原形, 比較級, 最上級) の文である。下記の表 1 にサンプル数, 総語数, 当該パターン数を示す。

表 1: 分析対象のサンプル数, 総語数, 当該パターン数

		A2	B1	B2+
Written	sample	154	228	18
	token	68,529	101,981	8,532
	BE DT/NN*/JJ*	1,440	1,983	166
Spoken	sample	30	90	30
	token	7,211	24,610	9,916
	BE DT/NN*/JJ*	154	419	183

上述の表が示すように、抽出されたパターンの最小数が 154 であったため、ランダム関数を用いて各データから 150 例ずつ無作為にサンプリングを行った。その抽出された書き言葉と話し言葉の A2, B1, B2+ の各 150 例 (計 900 例) をマニュアルでうなぎ文か否かの判定を行った。

うなぎ文の判定は、まず研究チームの 2 名が個別に判定を行い、次に研究チームの 4 名全員の合議で一致した例のみをうなぎ文と認定した。最初の 2 名による個別の判定の一致度は書き言葉は 30-83%, 話し言葉は 63-100% であった。話し言葉における言い直し (self-repair) のケースは、カウントしないことにした。たとえば次のケースにおいて、太字の文はうなぎ文のようにみえるが、次のイタリックの文で言い直している。そのような文は除外することにした。“And we – : before eating something , **our tongue is change**. Our – : sorry, *our tongue feel not at all feels appropriate* – : food, food.” (S B1)

4. 結果

4.1 量的結果

表 2: ランダム抽出した 150 ケース中の「うなぎ文」の数と割合

		A2	B1	B2+
Written	n	3	7	4
	%	2.00	4.67	2.67
Spoken	n	4	6	2
	%	2.67	4.00	1.33

表 2 にランダム抽出した 150 ケース中で「うなぎ文」と判断された文数と全体の割合を示す。結果は、言語のモード (書き言葉／話し言葉), 能力層 (A2, B1, B2+) に関わらず、うなぎ文の数は非

常に限られていた。各ケースが非常に少ないため、X² 乗検定等の統計処理を行っていないが、差なしと言ってよい。150 文の名詞／形容詞述語文の内、1.33-4.67%の発生率であった。後述するが、初級の A2 の話し言葉と上級の B2+の書き言葉の発生率が同じであることは着目に値する。

4.2 質的結果

現在の分析のところ、話し言葉、書き言葉、能力層において大きな違いは確認していない。

Teaching is high salary. (W A2) A part time job is a place for meeting. (W B1)

and second reason is that **part time job is a good situation** to learn – : learn relationship with other person such like boss or co-workers or some other persons (S B2)

5. まとめ

本研究では、日本人英語学習者の英語における 1) 名詞述語文と形容詞述語文に対応する「うなぎ文」の使用、および 2) 能力の伸長によるその使用傾向の変化について分析した。その結果、話し言葉であれ書き言葉であれ、うなぎ文の発生頻度はさほど高くなく、能力が伸長してもさほど変化はないことが判明した。本結果をふまえた現時点での解釈、制約および課題について言及する。

5.1 解釈

本結果は第二言語習得や英語教育の観点からは朗報かもしれない。主語と述語の一致関係が薄く、その意味の透明性がコンテキストに依存すると言われるうなぎ文はよく問題視されてきたものの、その発生率は大学生の A2 レベルでもさほど高くないようである。本結果からは、上述の「L1 において多用される表現が必ずしも L2 で多用されるわけではない」ケースと言える。付け加えて A2 のスピーキングと B2 のライティングのうなぎ文発生率が同じなのは注目に値する。というのも、能力が低く、かつ時間的制約に縛られるスピーキングのパフォーマンスでは、よりうなぎ文が発生しやすいと予測されるからである。したがって指導において、さほど労力をかけなくてもよいかもしれない。

一方、国際英語論(特に World Englishes)の観点から本結果をみると、僅かな使用頻度とはいえ、「A2 から B2 へ能力が伸長しても堅調にその特徴が残る」とも解釈できる。割合はあくまで相対的なものであるので、他の英語話者と比べた際に 5%未満でも寧ろ「高い」と言えるのかもしれない。B2 のライティングにおいても、“...they can learn what is important to make relations with colleagues or bosses better. In my case, **I am English** for high school students in a cram school as a part time job...” という典型的な「うなぎ文」を確認することもできる。

5.2 制約と課題

比較的大規模な ICNALE という学習者コーパスを用いたが、確認されたケースが少ないゆえに解釈が難しい。今回は名詞述語文と形容詞述語文の限られたパタンのみが分析対象である。またデータも日本人大学生のみしか分析していない。ICNALE が所収する「うなぎ文」を認めると言わ

れる韓国語母語話者データや同種の文を比較的許さない主語述語型言語の母語話者データと比較が必要である。またより低い能力層と推定される中学生・高校生の書き言葉／話し言葉コーパス (e.g., JEFL, LOCSE) などと同様の分析をすることが次の課題だろう。

¹ 事態「描写」の型の違いについての記述は数多くあるが、事態「把握」のあり方、特にその過程を明らかにした研究は管見の限りほとんど見当たらない。

² 料理の注文の際に「僕はうなぎ (だ)」でも通じるという現象に基づいて「うなぎ文」と呼ばれるようになったとされる (文の詳細な研究は奥津 (1978) を参照)。

謝辞

本研究は JSPS 科研費 20K00794, 「日本語母語話者の事態描写の「型」に即した「学習にやさしい英語」の研究」の助成を受けたものです。

引用文献

- Boroditsky, L. (2001) Does language shape thought? : Mandarin and English speakers' conceptions of time. *Cognitive Psychology* 43,1-22.
- D'Angelo, J., Yamaguchi, T., & Fujiwara, Y. (2021). Features of Japanese English. In Ee Ling Low & Anne Pakir (Eds.), *English in East and South Asia: Policy, features and language in use* (pp. 122-136). Routledge.
- Hinds, J. (1986). *Situation vs. person focus*. 西光義弘 (注). くろしお出版.
- Hino, N. (2021). Japanese English as an Expanding Circle variety: viewpoints and approaches. *Asian Englishes*, DOI: 10.1080/13488678.2020.1858582
- Honna, N. (2010). That restaurant is delicious [Japan]. *Asian Englishes*, 13(2), 64-65.
- 池上嘉彦 (1981) 『「する」と「なる」の言語学: 言語と文化のタイポロジーへの試論』大修館書店.
- 今井むつみ (2020) 『英語独習法』岩波書店.
- Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In: S. Ishikawa (Ed.), *Learner Corpus Studies in Asia and the World* (Vol. 1, pp. 91-118). Kobe University.
- 伊藤創 (2018) 「日英語母語話者における事態の描き方の型の違いと事態の捉え方の型の違い」『言語研究』154, 153-175.
- 伊藤創・藤原康弘 (2021) 「高校英語教科書における自動詞構文からみる学習にやさしい英語の考察」『JACET 中部支部紀要』19, 37-59.
- 伊藤創・岩男考哲・藤原康弘・仲潔 (2022a) 「日本語学から考える「日本英語」とその教育的可能性」『日本「アジア英語」学会第 49 回全国大会予稿集』pp. 26-27.
- 伊藤創・岩男考哲・藤原康弘・仲潔 (2022b) 「学習者の母語に即した「やさしい言語 (日本語) 習得」をめざして」*Proceedings of the 6th Asia Future Conference*, 705-714.
- 木村哲也 (1993) 『英語らしさに迫る-日本語の発想・英語の視点』研究社.
- 小宮富子 (2016) 「日本人英語の文法」塩澤正・吉川寛・倉橋洋子・小宮富子・下内充 (編) 『「国際英語論」で変わる日本の英語教育』(pp. 119-152). くろしお出版.
- 国広哲弥 (1974) 「日英語表現体系の比較」『言語生活』270, 46-52.
- 益岡隆志・田窪行則 (1992) 『基礎日本語文法一改訂版一』くろしお出版.
- May, S. (2013). *The multilingual turn*. Routledge.
- 奥津 敬一郎 (1978) 『「ボクハウナギダ」の文法』くろしお出版.
- Sasaki, M. (1990). Topic prominence in Japanese EFL students' existential constructions. *Language Learning*, 40(3), 337-368.
- Shibata, M. (2006). Topic making in English composition by Japanese EFL learners. *Scriptsimus*, 15, 1-26.
- Slobin, D. (1987). Thinking for Speaking. *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, 435-445.
- 外山滋比古 (1973) 『日本語の論理』中央公論社.
- 外山滋比古 (1992) 『英語の発想・日本語の発想』NHK 出版.
- 橘広司 (2014) 「諸英語に学ぶ〈日本英語〉実践化の試み」『英語教育』62(13), 60-62.

英作文における動詞・名詞コロケーション産出に対する DDL の効果

佐竹 由帆(青山学院大学)
t31330@aoyamagakuin.jp

The Effects of DDL on the Production of Verb-Noun Collocations in English Essays

SATAKE Yoshiho (Aoyama Gakuin University)

Abstract

Knowledge of collocations is important, while accurate use of collocations is difficult for English language learners. The author's previous study (Satake 2021) suggested that data-driven learning (DDL), in which learners learn by referring to a corpus, is effective for memorizing verb-noun collocations. Therefore, this study examined the effects of DDL on the production of verb-noun collocations in English composition. The participants were 19 first-year Japanese university students who were intermediate to upper-intermediate learners of English. They studied two verb-noun collocations by searching for their examples in the Wordbanks Online corpus once a week for nine weeks. The results showed that the target collocations, especially the collocations that the participants already knew, were used more in post-writing than in pre-writing, suggesting the effectiveness of DDL in producing collocations.

Keywords

データ駆動型学習(DDL), コロケーション産出, 動詞・名詞コロケーション, L2 ライティング

1. はじめに

コロケーションは慣習的に共起して用いられる 2 語以上の語句で (Firth, 1957), 多くの単語はコロケーションの形で使用されるため外国語教育においてコロケーションの知識は重要だが, 英語学習者にとってコロケーションを正確に使用することは難しい (e.g., McIntosh et al., 2009)。筆者の先行研究 (佐竹, 2021) はコーパスを参照して学習するデータ駆動型学習 (Data-driven learning: DDL) が動詞・名詞コロケーションを記憶する上で有効であることを示唆したため, 本研究は英作文における DDL の動詞・名詞コロケーション産出に対する効果を検証する。

2. 先行研究

英語学習者にとってコロケーションの正確な使用は難しい。Laufer and Waldman (2011) はルーベン大学英語母語話者コーパスの 220 の頻出名詞について動詞・名詞コロケーションを抽出

してヘブライ語母語話者が書いた英文の動詞・名詞コロケーションと比較し、英語学習者は英語母語話者よりコロケーション使用が少なく、上級レベルになっても誤りが残ることを示した。

筆者の先行研究、佐竹(2021)は、日本の大学生を対象に DDL の動詞・名詞コロケーション記憶に対する効果を検証した。参加者は 19 名の大学生で、英語中級・中上級の学習者だった。ワークシートに書かれた二つの間違っただ動詞・名詞コロケーションをコーパス参照して修正するタスクを週 1 回 9 週間行い事前事後テストで効果検証した結果、参加者の点数は事後テストで有意に向上し効果量大だった。DDL は動詞・名詞コロケーションの記憶を促進することが示唆されたが、日本の英語学習者を対象とする DDL の動詞・名詞コロケーション学習に対する効果については先行研究が少なく、産出に対する効果はまだ不明である。ゆえに、本研究のリサーチクエスションは、コーパスで動詞名詞コロケーションを参照することはコロケーション産出を促進するか、とする。

3. リサーチデザイン

3.1 参加者

参加者は(佐竹, 2021)と同じで、2021 年後期に筆者が担当した英語ライティングの授業を必修で受講した 19 名の大学生である。22 名の受講生のうち、本研究についての説明後書面で研究参加の同意を取れた 19 名を対象とし、事前事後ライティング及び事前事後遅延事後テストすべてに参加した 18 名のデータを使用した。学生は 12 週間本研究に参加した。参加者の英語力は、ヨーロッパ言語共通参照枠(CEFR)の B1 から B2 だった。

3.2 タスク

参加者は、学習対象コロケーションと比較して使用頻度の低い動詞・名詞の組み合わせを含む英文二文と訳(例: “Madame Curie could **destroy** the mold as the first female Nobel-Prize winner. キュリー夫人は女性初のノーベル賞受賞者として殻を破ることができました。”)が書かれたワークシートを配布され、ワードバンクスオンラインを参照して 15 分以下下線の名詞と共起する動詞を調べ、用例のコンコーダンスラインを参照してより自然な英文を書き、参照した用例を転記し、修正の理由を書いて提出した(第 2-10 週)。課題英文の修正対象の動詞は、和訳の動詞に対応する意味だが課題英文の名詞との組み合わせでは使用頻度が低いものを選んだ。フィードバックとして、ワークシート回収直後に英英辞書(主に Cambridge, Collins, Merriam-Webster, 2022)で対象コロケーションのページを筆者がクラス全体に提示し、短い明示的解説をした。他の学習方法の影響を排除するため、学生は授業中に他の方法で対象コロケーションを学習することはなかった。

本研究の目的は学習した動詞・名詞コロケーションの産出に対する DDL の効果であり、学習後に書く英語エッセイで使用可能なコロケーションを学習対象として選ぶ必要があるため、対象コロケーションは、使用テキスト(Pike-Baky & Blass, 2006)の語彙学習ページに掲載されているものを中心に、事前事後のエッセイ(3.4 参照)に使用できそうなものを 18 選んだ(事前事後テスト問題番号順に、1)“provide (a/the) service(s)”, 2)“go bankrupt”,

3) “turn/make a profit”, 4) “make a/one’s pitch”, 5) “take/run a risk”, 6) “break the mold”, 7) “give (a) direction(s)”, 8) “push (a/the) boundary/boundaries”, 9) “bear a/the burden”, 10) “cast (a) vision”, 11) “require determination”, 12) “restore order”, 13) “take a chance”, 14) “stretch the/one’s limit(s)”, 15) “break a rule”, 16) “set the scene”, 17) “spin one’s wheels”, 18) “cast doubt”。

3.3 参照コーパス

参加者が使用したコーパスは、英・米・豪・加などの英語について様々なジャンルの書き言葉・話し言葉約 6 億語を収録した大規模コーパス、ワードバンクオンラインである (NetAdvance, 2003)。ワードバンクオンラインを参加者が容易に使用できるように、単語やフレーズを検索するための使いやすいインターフェイスをオンラインで提供している小学館コーパスネットワークを使用した。参加者は前期にコーパスを使用した経験があったが、前期とは使用コーパスが異なるため、最初の修正タスクの前に参加者に使用法について 10 分間指導した。参加者は単語を検索してその用法に関する情報を得たり、コンコーダンスラインを適切に解釈したりする方法を学習した。

3.4 エッセイ

タスクのコロケーション産出の効果を検証するため、対象コロケーションの事前事後エッセイにおける産出数を比較した。参加者は第 1 回目のタスクの前週に事前エッセイを (第 1 週)、最終修正タスクの翌週に事後エッセイ (第 12 週)、“**success in business**”という同一のトピックについてそれぞれ 30 分で行った。筆者はエッセイを収集して事前エッセイと事後エッセイの二つの学習者コーパスを作成し、使用された対象コロケーション (3.2 参照) の数を比較した。

3.5 テスト

タスクの効果検証の参考にするため、対象コロケーションの事前事後エッセイの産出数比較に加えて、産出事前・事後・遅延事後テストを行った。事前テストは第 1 回目のタスクの前週に全問について行った (第 1 週)。事後テストは各問についてタスク実施からテストまでの時間をできるだけ同じにして条件を揃えるため、毎回タスクの次週に二つずつ行った (第 3-11 週)。遅延事後テストは、最終事後テストの翌週に全問について行った (第 12 週)。事後テストのみ二問ずつ出題で解答時間 30 秒、事前・遅延事後テストは全問出題で解答時間 5 分と出題形式は異なるが、問題は同一のものを使用した。英語名詞と和訳を見て適切な動詞を書き入れる問題で (例、“() the mold 型を破る”)、タスクで学習した 18 の動詞・名詞コロケーションを出題した。正誤については、学習対象のコロケーション (3.2 参照) でなくても、同意かつワードバンクオンラインで当該名詞とのコロケーションで一定数以上の頻度がある動詞については正答とし、1 問 1 点、満点 18 点で採点した。

3.6 手順

本研究の手順は下記の通りである。

- 1) 事前エッセイ (30 分, 第 1 週)
- 2) 事前テスト (5 分, 第 1 週)
- 3) ワードバンクオンラインの使用法指導 (10 分, 第 2 週)
- 4) 動詞・名詞コロケーション修正タスク 2 題 (15 分, 第 2-10 週)
- 5) フィードバック (1・2 分, 第 2-10 週)
- 6) 事後テスト (30 秒, 第 3-11 週, タスク・フィードバックの翌週)
- 7) 事後エッセイ (30 分, 第 12 週)
- 8) 遅延事後テスト (5 分, 第 12 週)
- 9) 分析

事前・事後エッセイの対象コロケーションの産出数の差の有効性検証のために、ウィルコクソンの符号順位検定を行った。ノンパラメトリック検定を使用したのは、データの正規性の前提が満たされなかったためである。テスト間の点数の差の検証のために一元配置分散分析を行った。効果量の判断は、第二言語に関する研究における基準を提案した Plonsky and Oswald (2014) による。質的分析のためには、参加者が提出したワークシートを参照した。

4. 結果と考察

4.1 事前事後エッセイにおける対象コロケーション産出

図 1 事前事後エッセイにおける対象コロケーション産出

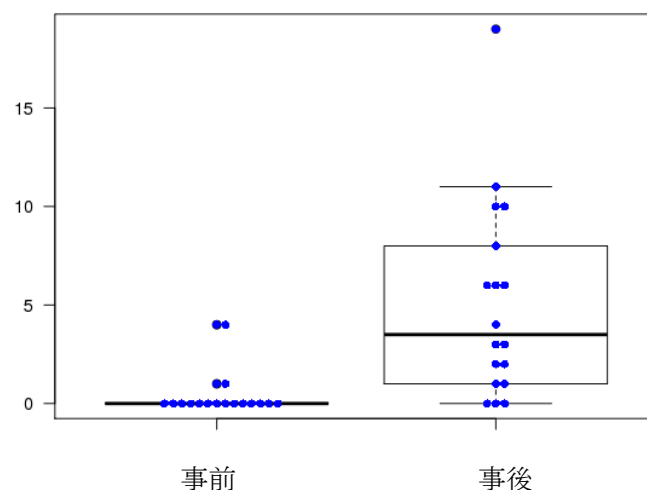


図 1 の事前事後エッセイにおける対象コロケーション産出が示すように、参加者のコロケーション産出数は事後エッセイで事前エッセイよりも増加した。事前テストの平均産出数 0.56, 事後エッセ

イの平均産出数は 5.11 点で事後エッセイの方が高く、ウィルコクソンの符号順位検定の結果は有意差ありで効果量は大に近かった($z=3.41$, $p<.001$, $d=1.38$)。コロケーション別に見ると、テストの問題番号別の産出を表す表 1 が示すように、事後テストで産出数が向上したのが 18 中 15、未使用のまま変わらなかったのが 3 で、大多数で向上が見られた。ゆえにタスクは動詞-名詞コロケーションを産出する上で有効で、全体的に事後エッセイにおける産出数が上昇したと言える。

表 1 事前事後エッセイにおける対象コロケーション別産出数

問題番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	計
事前	4	1	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
事後	10	6	19	3	11	6	8	1	4	10	3	2	6	0	0	0	1	2	92

4.2 事前・事後・遅延事後テストの点数比較

図 2 事前・事後・遅延事後テストの平均点

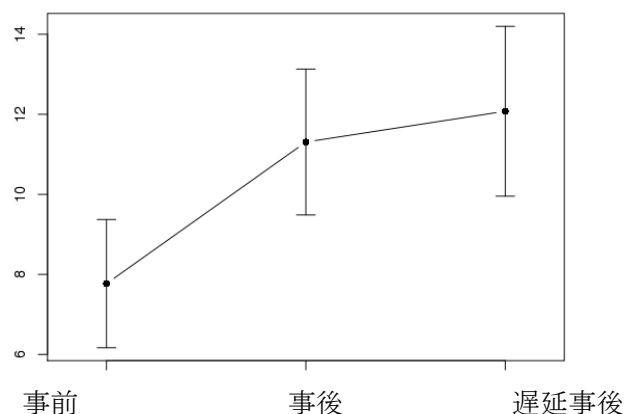


表 2 事前事後遅延事後テストにおける対象コロケーション別点数

問題番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	計
事前	12	15	18	3	17	13	6	1	7	9	7	1	8	4	14	5	0	1	141
事後	13	15	18	8	17	15	12	2	13	14	10	7	12	9	15	14	5	5	204
遅延事後	13	15	18	10	17	16	12	3	13	16	10	9	13	9	16	15	6	6	217

図 2 の事前・事後・遅延事後テストの平均点のグラフが示すように、参加者のテストの点数は事後・遅延事後テストで事前テストよりも増加した。事前テストの平均点は 18 点満点中 7.8 点、事後テストの平均点は 11.3 点で遅延事後テストの平均点は 12.1 点で、分散分析の結果テスト間に有意差があった($F(2, 34)=7.60$, $p=.007$, $\eta_p^2=.39$)。下位検定の結果、事後テスト及び遅延事後テストの点数と事前テストの点数間に有意差があるが(事前・事後間: $p=.006$, 事前・遅延事後間: $p<.001$)、事後テストと遅延事後テスト間に有意差はなかった。

($p=.61$)。コロケーション別に見ると、表 2 が示すように、事後・遅延事後テストで点が向上したのが 18 中 15、変わらなかったのが 3 で、大多数で向上が見られた。ゆえにタスクは動詞 - 名詞コロケーションを記憶する上で有効であったと言える。

4.3 考察

表 1 と表 2 を比較すると、コロケーション産出が事後に 5 回以上増えたのは 8 つだが (1,2,3,5,6,7,10,13), そのうち事後・遅延事後テストで 5 点以上向上したのは 2 つのみであった(7,10)。DDL タスクは、タスクにより新たに記憶したコロケーションより、既知のコロケーションの産出を促進したと考えられる。DDL は未知のコロケーションの記憶に貢献するだけでなく、既知だが産出語彙となっていないコロケーションの産出を促進する効果があることが示唆された。

5. 終わりに

本研究は英作文における DDL の動詞-名詞コロケーション産出に対する効果を検証し、DDL が未知のコロケーションの記憶・既知のコロケーションの産出を促進することを示し、既知の語彙を産出語彙にするための学習活動としての DDL の有効性が示唆された。今後の課題としては、参加者の人数を増やして研究規模を大きくし、対照群と比較検証したい。

謝辞

本研究は JSPS 科研費 JP20K13109 の助成を受けたものです。

引用文献

- Cambridge dictionary online. (2022). Cambridge University Press.
<https://dictionary.cambridge.org/>
- Collins online dictionary (2022). Collins. <https://www.collinsdictionary.com/>
- Laufer, B. & Waldman, T. (2011). Verb-noun collocation in second language writing: a corpus analysis of learners' English. *Language learning*, 61(2), 647-672.
- Merriam-Webster online. (2022). Merriam Webster. <https://www.merriam-webster.com/>
- NetAdvance Inc. (2003). Shogakukan corpus network.
<https://scnweb.japanknowledge.com/>
- Pike-Baky, M. and Blass, L. (2006). *Mosaic 1: Writing* (5th ed.). McGraw-Hill Education.
- Plonsky, L., and Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language learning*, 64(4), 878-912.
- 佐竹由帆 (2021) 「英語の動詞-名詞コロケーション学習に対する DDL の効果」『英語コーパス学会大会予稿集 2021』, 85-90.

A Critical Evaluation of the Optimal Association Measures for Creating L2 Learners' Collocations Lists

Fukuda Kohei (Tokyo University of Foreign Studies, Graduate Student)

fukuda.kohei.r0@tufs.ac.jp

Tono Yukio (Tokyo University of Foreign Studies)

y.tono@tufs.ac.jp

Abstract

Association measures (AM) have been used to extract candidates for the collocations list for L2 learners. However, little attention has been paid to which AM is optimal for extracting collocations for pedagogical purposes. Furthermore, dispersion measures are often not considered for the evaluation of AMs. This study explores which AM is suitable for identifying pedagogically useful English collocations, considering both frequency and dispersion thresholds. Adjective-Noun, Verb-Object, Adverb-Adjective and Adverb-Verb pairs were extracted from the syntactically parsed British National Corpus, and their AMs (MI, MI2, MI3, t-score, z-score, logDice, log-likelihood) were calculated. Dispersion was calculated using Gries' DP. AMs were evaluated against the three gold standards, and the precision and recall were computed by setting varying frequency and DP thresholds. The results shows that it is extremely important to select an optimal AM according to the syntactic structures underlying collocations, with t-score or raw frequency being optimal for Adjective-Noun and Verb-Object pairs, logDice or MI3 for Adverb-Adjective pairs, and log-likelihood for Adverb-Verb pairs. Applying frequency and DP thresholds at an appropriate level can be effective. While an appropriate frequency threshold can vary, an adequate DP threshold is mostly 0.6 or 0.7.

Keywords

corpus linguistics, collocation, association measure, dispersion, SLA

1. Introduction

It is widely acknowledged that collocational knowledge is an important aspect of second language acquisition. To facilitate collocation learning, it is essential to identify collocations of pedagogical use, but there is no consensus on which collocation should be learned. Therefore, we need to first create a collocations list to which learners, teachers and material developers can refer.

Although there are several collocations dictionaries for learners and collocations lists

for EAP, few attempts have been made to create a collocations list for general pedagogical purposes with elementary and intermediate levels in mind. As a first step towards creating a collocations list, this study explores a method that automatically identifies pedagogically relevant collocations from corpora using association measures (AM).

2. Literature Review

Although quite a few association measures have been proposed, there are no AMs that can be universally used, so it is crucial to select a suitable AM depending on the aim of the research (Evert, 2008). Gablasova et al. (2017) argue that for proper use of AMs, we need to understand their mathematical reasoning, the scale on which they operate and their practical effect. They question the dominance of t-score and MI — particularly MI — in language learning research (LLR), and suggest that we should also consider the use of logDice or MI2 instead. Furthermore, they claim that dispersion should be taken into consideration “because collocations that are more general (i.e., occur across a variety of contexts) are more likely to be encountered by language users regardless of the context of use” (p.160).

It has been shown that extracting collocations based on syntactic dependency from a syntactically parsed corpus is a better method than window span, which has been widely used in LLR in corpus linguistics (Bartsch & Evert, 2014). Bhalla & Klimcikova (2019) argue that syntactic structures underlying collocations significantly impact the results. Thus, it is important to assess the performances of AMs, considering different syntactic structures.

3. Research Design

3.1 Aim of the Research

The present study focuses on AMs’ practical effect because the primary aim is the practical application of AMs to creating a collocations list. Taking the studies mentioned above into account, this study attempts to identify an optimal parameter in terms of AMs and frequency and dispersion thresholds to extract pedagogically useful collocations from corpora, with the differences between syntactic structures in mind. Furthermore, this research explores whether there are any differences between the optimal parameters for extracting collocations for pedagogical purposes and those not limited to a pedagogical context. AMs to be investigated include MI, MI2, MI3, t-score, z-score, logDice and log-likelihood as well as raw frequency. We used DP (Deviation of Proportions) as a dispersion measure (Gries, 2008).

3.2 Method

The authors syntactically parsed the BNC and extracted word pairs of five syntactic patterns (adjective + noun [*amod*], noun + noun [*nounmod*], verb + object [*obj*], adverb + adjective [*advmod-adj*], adverb + verb [*advmod-verb*]) on syntactic dependency. Then, we calculated association measures and DP for all the word pairs extracted. Because the present research is interested only in positive associations, word pairs whose observed frequencies are lower than their expected frequencies were excluded, leaving 3,016,262 word pairs (see table 1).

To find an optimal parameter, we adopted a gold standard method, an oft-used methodology for the comparative evaluation of association measures in the automatic extraction of collocations. In this approach, performances of association measures are assessed against a gold standard (a supposed set of true collocations) in terms of precision and recall of *n* highest-ranked list extracted using different AMs (n-best list).

Since different gold standards can yield different results, we used three sets of gold standards. Two of them are intended as a gold standard of pedagogically useful collocations. The other is intended as a set of true collocations that are not limited to a pedagogical context. All the three gold standards used lexical collocations (*amod*, *obj*, *advmod-adj*, *advmod-verb*) listed in Oxford Collocations Dictionary (OCD). This study treats *amod* and *nounmod* as one syntactic structure (*amod*) because OCD does not regard them separately.

Table 2 shows the description of the three sets of gold standards. The first (GSL-GS) used New General Service List (new-GSL), which contains a stable core general vocabulary identified by comparing four language corpora (Brezina & Gablasova, 2015). The second (CB-GS) used a corpus compiled from English course books¹. The third (OCD-GS) comprises all the collocations of targeted syntactic patterns listed in OCD. Word pairs which did not occur in BNC were excluded from each of the gold standards. Performances of AMs are evaluated against the three gold standards in terms of

Table 1 *The number of word pairs*

syntactic structures	number
adjective + noun	1,042,621
noun + noun	581,829
verb + object	984,642
adverb + adjective	127,670
adverb + verb	279,500
total	3,016,262

Table2 *Gold standard*

Gold standard	description	syntactic patterns	number
GSL-GS	collocations listed in OCD which are composed only of words that are on New General Service List.	<i>amod</i>	23,737
		<i>obj</i>	12,359
		<i>advmod-adj</i>	1,370
		<i>advmod-verb</i>	2,895
CB-GS	collocations that occur in the course books and listed in OCD	<i>amod</i>	7,081
		<i>obj</i>	5,708
		<i>advmod-adj</i>	717
		<i>advmod-verb</i>	1,239
OCD-GS	collocation that are listed in OCD	<i>amod</i>	75,701
		<i>obj</i>	32,849
		<i>advmod-adj</i>	7,769
		<i>advmod-verb</i>	11,501

precision and recall, setting varying frequency threshold (from 0 to 50 increasing by 5) and DP threshold (from 1.0 to 0.1 decreasing by 0.1) (The closer an expressions' DP gets to 0, the more evenly distributed it is). Following Evert, Uhrig et al. (2017), we calculated average precision (AP) as a composite measure which serves as a summary of precision/recall. While the study used AP50 (average precision values up to 50% recall), we also calculated AP10 and AP30, considering the importance of narrowing down the number of collocations to include in a list for pedagogical use.

4. Results and Discussions

Table 3, 4, 5 and 6 show the parameters that returned the highest AP10/30/50 value against each of the three gold standards. Table 3 and 4 show that t-score or raw frequency with 0.6 or 0.7 DP threshold is an optimal parameter for *amod* and *obj*. Since frequency thresholds have little effect on the results of raw frequency and t-score, which is highly correlated with raw frequency, it is not necessary to apply frequency thresholds. Table 5 suggests that logDice or MI3 with a DP threshold of 0.6 or 0.7 seems to be an optimal parameter for *advmod-adj*. The performances of AMs other than t-score are more or less influenced by frequency thresholds, and therefore, applying an appropriate frequency threshold is desirable. However, it is difficult to determine the optimal frequency threshold since it varies depending on the gold standard and AP10/30/50. Setting it at minimum of 15 would improve extraction performances without

Table 3 *amod*

Gold standard	AP 10/30/50	AP value	Parameters		
			f tL	DP tL	AM
GSL-GS	AP10	59.8%	0~50	0.6	T
	AP30	51.4%	40	0.7	T
	AP50	43.0%	0	0.7	rf
CB-GS	AP10	49.4%	0~50	0.5	T
	AP30	35.8%	20	0.6	rf
	AP50	27.2%	20	0.6	rf
OCD-GS	AP10	73.8%	35	0.6	LL
	AP30	68.3%	10	0.7	LL
	AP50	60.0%	5	0.8	LL

Notes: f tL = frequency threshold

DP tL = DP threshold T = t-score

rf = raw frequency LL = log-likelihood

Table 4 *obj*

Gold standard	AP 10/30/50	AP value	Parameters		
			f tL	DP tL	AM
GSL-GS	AP10	80.0%	0~50	0.6	T
	AP30	64.7%	40	0.6	T
	AP50	52.2%	20	0.7	T
CB-GS	AP10	82.3%	0~50	0.5	T
	AP30	63.0%	35	0.6	rf
	AP50	48.5%	45	0.7	T
OCD-GS	AP10	83.5%	35	0.6	LL
	AP30	73.6%	10	0.8	LL
	AP50	62.2%	5	0.8	LL

Table 5 *advmod-adj*

Gold standard	AP 10/30/50	AP value	Parameters		
			f tL	DP tL	AM
GSL-GS	AP10	37.4%	50	0.5	D
	AP30	28.3%	25	0.6	D
	AP50	22.0%	15	0.7	D
CB-GS	AP10	39.8%	50	0.6	D
	AP30	29.2%	45	0.6	MI3
	AP50	23.3%	30	0.7	MI3
OCD-GS	AP10	67.3%	10	0.7	MI
	AP30	53.5%	5	0.9	D
	AP50	44.3%	5	0.9	D

Notes: D = logDice

Table 6 *advmod-verb*

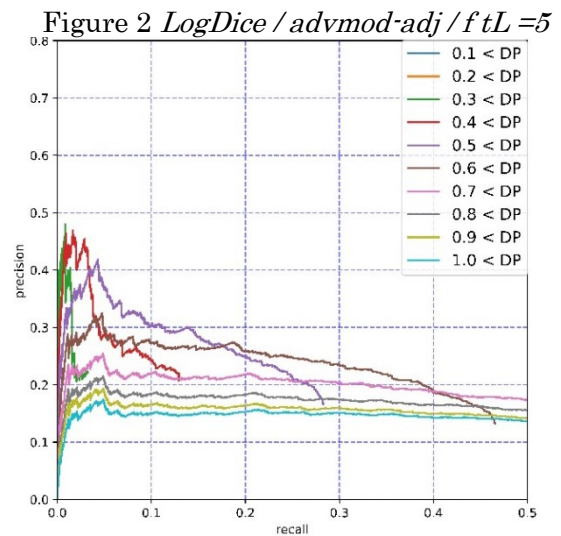
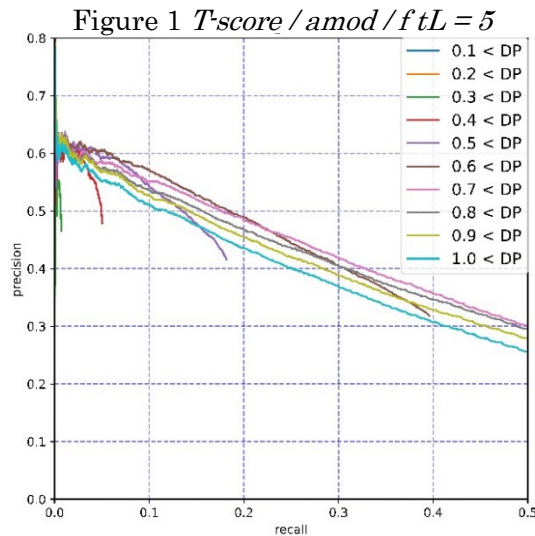
Gold standard	AP 10/30/50	AP value	Parameters		
			f tL	DP tL	AM
GSL-GS	AP10	29.3%	50	0.5	LL
	AP30	24.2%	25	0.6	LL
	AP50	20.5%	15	0.7	T
CB-GS	AP10	33.9%	50	0.4	Z
	AP30	28.8%	50	0.7	LL
	AP50	23.4%	40	0.8	LL
OCD-GS	AP10	63.2%	20	0.8	MI
	AP30	57.2%	10	0.9	MI
	AP50	50.6%	5	0.9	Z

Notes: Z = z-score

counterproductive effects, but if the number of collocations is to be narrowed down, a higher threshold could be set. According to Table 6, log-likelihood can be the optimal AM for extracting *advmod-verb* collocations. However, it is hard to decide the optimal threshold of frequency and DP. While it might vary on a case-by-case basis, DP thresholds of at least 0.7 and frequency thresholds of at least 25 would be appropriate. It should be noted that optimal association measures significantly vary depending on syntactic patterns. Therefore, selecting a suitable association measure and threshold according to the underlying syntactic structure is extremely important.

Furthermore, the results suggest that association measures appropriate for pedagogical purposes differ from those not limited to an educational context. In *amod* and *obj*, log-likelihood gives the highest performance when using OCD-GS, which is intended as a gold standard not limited to pedagogical use. In *advmod-adj* and *advmod-verb*, MI can be the most relevant association measure against OCD-GS; however it never comes at a high rank when using GSL-GS or CB-GS. Thus, we need to assess which AM works best within the desired context.

According to table 3, 4, 5 and 6, for GSL-GS and CB-GS, the best DP threshold ranges from 0.4 to 0.8, mostly 0.6 or 0.7. This suggests that setting DP thresholds at an appropriate level is effective. DP thresholds should not be lowered in excess; this may lead to adverse effects because it excludes collocations from the n-best list. Figure 1 is a precision/recall plot for t-score against GSL-GS setting varying DP thresholds, which indicates that as a DP threshold is lowered, average precision increases, but when a DP threshold is below 0.5, it impairs the performance. Figure 2 is a precision/recall plot for logDice against GSL-GS. The AMs highlighting exclusivity such as logDice are more greatly influenced by DP thresholds than t-score. As with figure 1, precision increases as a DP threshold is lowered, but when a DP threshold is below 0.4, it is clear that it



adversely affects the performance. Moreover, it should be recognized that DP thresholds limit the maximum value of recall. For instance, in figure 2, when a DP threshold is set at 0.5, the maximum value of recall is about 28%. It is important to set DP thresholds at an appropriate level, usually 0.6 or 0.7.

5. Conclusion

The present study suggests that it is extremely important to select a suitable association measure depending on syntactic structures. An AM which highlights frequency such as t-score or raw frequency itself is suitable for extracting *amod* or *obj* collocations. LogDice or MI3 can be an optimal AM for *advmod-adj*. Log-likelihood can be said to be suitable for *advmod-verb* collocations. These association measures can be improved by applying frequency thresholds and DP thresholds. To give a higher performance, it is necessary to apply frequency and DP thresholds at an appropriate level. While an optimal frequency threshold can vary depending on the situation, an adequate DP threshold is 0.6 or 0.7 in most cases.

Note

1. The CEFR Coursebook Corpus. An internal resource. TonoLab. Tokyo University of Foreign Studies. See Tono (2017) for a description of this corpus.

Bibliography

- Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*. 2nd Work Report of the Academic Network Internet Lexicography, OPAL—Online publizierte Arbeiten zur Linguistik. Mannheim: Institut für Deutsche Sprache.
- Bhalla, V., & Klimcikova, K. (2019). Evaluation of automatic collocation extraction methods for language learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 264–274.
- Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1–22.
- Evert, S. (2008). Corpora and collocations. In M. K. Anke Lüdeling (Ed.), *Corpus Linguistics. An International Handbook* (pp.1212–1248). Mouton de Gruyter.
- Evert, S., Uhrig, P., Bartsch, S., & Proisl, T. (2017). E-VIEW-affiliation—A large-scale evaluation study of association measures for collocation identification. In *Proceedings of eLex*, 531–549.
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, 67(S1), 155–179.
- Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437.
- Tono, Y. (2017). The CEFR-J and its impact on English language teaching in Japan. *JACET international convention selected papers*, 4, 31–52.

教科書コーパス分析における推定周辺平均値の有用性について —英語法助動詞の頻度と意味の分析から—

梶山 達也(同志社大学 大学院生)
tatsuyakajiyama@gmail.com

The usefulness of the estimated marginal means for the analysis of the textbook corpus —From a perspective of the frequencies and the meanings of English auxiliary verbs—

KAJIYAMA Tatsuya (Doshisha University, Graduate Student)

Abstract

This paper discusses the usefulness of the generalized linear mixed model (GLMM) and the estimated marginal mean (EMM) for the analysis of a small corpus. Japanese-English textbooks used in Japanese high schools were converted into text data. All English modal auxiliary verbs were extracted and classified into deontic/dynamic meanings and epistemic meanings. Then, the GLMM and the EMM were calculated. Results show that the frequencies of each meaning of each modal auxiliary verb increased as the grades advanced. This finding appears to suggest that the GLMM and the EMM be useful for the analysis of a small corpus.

Keywords

Modal Auxiliary Verb, Textbook Corpus, Generalized Linear Mixed Model, Estimated Marginal Mean

1. はじめに

本研究では、小規模コーパス分析における、一般化線形混合モデルと推定周辺平均値の有用性について考察する。しばしば、大規模コーパス分析では、カイ二乗検定やフィッシャーの正確確率検定といった、クロス集計表を使った検定が用いられる。その理由として、水本(2010, p. 10)は、「正確な p 値」と「コーパスにおけるサンプル(標本)と母集団の関係」をあげている。ここで言う「正確な p 値」とは、 p 値の計算に「母集団の未知のパラメータや、サンプリング誤差(sampling error)」(水本, 2010, p. 6)を含めずに計算された値である。言い換えると、「正確な p 値」とは、収集されたデータ(サンプル)が考察対象の母集団と等しい場合に得られる p 値を指す。例えば、1 人の作家の作品における、ある単語の頻度を調べるためには、作家の全ての作品をコーパス化すれば、母集団に対して統計検定を行うことができる。データの特性を考慮すれば、適切な手法であろう。し

かし、規模の小さいコーパスを分析する場合は、統計結果が歪曲する可能性があり注意が必要である。ここでは、先行研究を概観した後で、教科書コーパスを用いて、小規模コーパス分析における一般化線形混合モデルと推定周辺平均値の有用性を考察する。

2. 先行研究

2.1 コーパス分析における統計使用の現状

水本(2010)では、サンプルサイズが小さい場合の、パラメトリック検定とノンパラメトリック検定、正確な p 値を得られる並び替え検定や直接確率検定を比べ、それぞれの検出力を比較している。そして、「正しい p 値が確認することができるという点で、並び替え検定・直接確率検定による方法はまず初めに推奨されるべき」(水本, 2010, p. 13)と指摘している。

また、Newbery-Payton(2022)は、中国人英語学習者と日本人英語学習者の“will”の使用頻度について、エッセイデータを元に比較している。このデータは、the International Corpus Network of Asian Learners of English(ICNALE)(Ishikawa, 2013)から抽出されており、part-time job(PRJ)と smoking(SMK)の話題について書かれたエッセイである。Newbery-Payton は、クラスカル・ウォリス検定を行い、SMK データで、有意差があったと報告している。

さらに、梶山(2022.3)では、中学校・高等学校英語検定教科書で用いられる、英語法助動詞の各意味の頻度について、カイ二乗検定とフリードマン検定を行い、学年別と法助動詞別に有意な差があったと指摘している。

このように、コーパス分析では、データの特性に合わせて、クロス集計表を用いた分析やノンパラメトリック検定が使われている。しかし、これらの手法に関して次のような問題点が指摘されている。

2.2 クロス集計表やノンパラメトリック検定の限界

ノンパラメトリック検定は順位データを用いるが、Brezina(2018, p. 195)によると、“work with ranks; relative frequencies (scale variable) can easily be transformed into ranks, although this is at the cost of losing some information about the exact differences between speakers (cases)”とされる。つまり、ノンパラメトリック検定の場合、得られた得点や頻度データを順位データに置き換えており、その過程で、情報力が落ちてしまう。しかし、データ分析において、可能な限り、情報量を保持したまま分析を行うべきである。

また、森(2020)は、均衡コーパスのように大規模なコーパスで用いる手法を、学習者コーパスのような小規模コーパスに用いる危うさを指摘している。その理由として、「独立性の仮定を満たしていないことや、学習者の頻度を合計した場合、外れ値によって全体の傾向性が大きく歪む可能性」をあげている。(森, 2020, p. 444)。

さらに、これらの統計手法は、2 つ以上変数を持つデータを処理する場合、交互作用を確認することができない。この欠点を補う手法として、一般化線形混合モデルと推定周辺平均値(Searle et al., 1980)による分析について考察する。

3. リサーチデザイン

3.1 研究目的と研究課題

本発表の目的は、小規模コーパス分析における一般化線形混合モデルと推定周辺平均値の有用性を確かめることである。そこで、次のような研究課題を立てる。

RQ: 小規模コーパス分析において、一般化線形混合モデルと推定周辺平均値は、有用な手法となり得るだろうか。

以下では、本発表で使うデータや手法について詳しく述べる。

3.2 データ

本発表では、高等学校英語検定教科書（2012 年～2014 年検定）で使用されている、英語法助動詞の各意味の頻度に注目する。これらの教科書の本文を全てテキストデータ化し、CasualConc (Version 2.1.7) (Imao, 2021) を使い、頻度が極端に少なかった“shall”を除く、8 つの法助動詞の用例を集めた。その後、意味分類の基準を複数の評価者で作り、この基準に従って、全ての用例の意味を分類した。その結果を表 1 に記す。

表 1 各法助動詞の各意味の頻度

法助動詞	意味	Grade 1	Grade 2	Grade 3	法助動詞	意味	Grade 1	Grade 2	Grade 3
can	dynm	550	770	739	could	dynm	247	353	418
	deon	22	46	40		deon	12	21	20
	epst	46	96	110		epst	22	51	74
may	deon	11	10	17	might	deon	1	0	3
	epst	155	211	286		epst	48	128	165
should	deon	108	142	161	must	deon	72	81	99
	epst	21	42	50		epst	36	51	51
will	dynm	114	128	104	would	dynm	93	137	138
	epst	377	533	502		epst	174	405	461

Note. “dynm” = dynamic meaning, “deon” = deontic meaning, “epst” = epistemic meaning

3.3 手法

本発表では、一般化線形混合モデルを使って分析をする。応答変数である各意味の頻度は、上限なしのカウントデータであり、平均と分散が等しくないため、確率分布に負の二項分布を指定しモデル推定を行った。個体差があると考え、教科書をランダム効果にした。5 つのモデルを仮定し、最適なモデルを法助動詞ごとに推定した。表 2 に、can を例に挙げ、探索したモデルを列挙する。

表 2 と同じように、他の法助動詞もモデル推定を行い、Akaike’s Information Criterion (AIC) が最も小さいモデルを採用した。各法助動詞の最適モデル結果を表 3 に掲載する。その後、推定周辺平均値を求め、各意味と各学年のどこに有意差があるかを分析した。

表 2 can の意味頻度についての最適モデル探索

モデル	説明変数	AIC	χ^2 値	自由度	p 値
null.model	切片	1484.9	—	—	—
model 1a	意味	1146.0	342.97	2	< 0.01
model 1b	学年	1484.6	0.00	0	—
model 2a	意味 + 学年 (交互作用なし)	1111.8	376.81	2	< 0.01
model 2b	意味 + 学年 (交互作用あり)	1109.2	10.61	4	< 0.05

表 3 最適モデルの結果

法助動詞	説明変数	予測値	標準誤差	z 値	p 値	法助動詞	説明変数	予測値	標準誤差	z 値	p 値
can	Intercept	-0.085	0.224	-0.379	0.705	could	Intercept	-0.829	0.194	-4.267	<.01
	usedynm	3.170	0.216	14.693	<.01		usedynm	2.918	0.149	19.534	<.01
	useepst	0.700	0.254	2.755	<.01		useepst	0.931	0.168	5.548	<.01
	grade 2	0.737	0.253	2.907	<.01		grade 2	0.471	0.109	4.322	<.01
	grade 3	0.515	0.265	1.941	0.523		grade 3	0.693	0.110	6.309	<.01
	usedynm:grade 2	-0.410	0.268	-1.530	0.126	may	Intercept	-1.135	0.229	-4.946	<.01
	useepst:grade 2	-0.028	0.314	-0.090	0.928		useepst	2.837	0.176	16.113	<.01
	usedynm:grade 3	-0.204	0.280	-0.731	0.465		grade 2	0.266	0.138	1.926	0.054
	useepst:grade 3	0.383	0.322	1.189	0.235		grade 3	0.649	0.137	4.742	<.01
will	Intercept	1.484	0.150	9.879	<.01	might	Intercept	-4.047	0.563	-7.192	<.01
	useepst	1.127	0.150	7.522	<.01		useepst	4.459	0.516	8.643	<.01
	grade 2	0.243	0.164	1.488	0.137		grade 2	0.949	0.236	4.016	<.01
	grade 3	-0.071	0.175	-0.403	0.687		grade 3	1.348	0.238	5.659	<.01
	useepst:grade 2	0.142	0.206	0.688	0.491	must	Intercept	0.918	0.170	5.402	<.01
would	useepst:grade 3	0.461	0.216	2.132	<.05		useepst	-0.594	0.118	-5.040	<.01
	Intercept	1.086	0.195	5.583	<.01		grade 2	0.187	0.145	1.293	0.196
	useepst	0.604	0.171	3.542	<.01	should	grade 3	0.378	0.145	2.606	<.01
	grade 2	0.470	0.176	2.675	<.01		Intercept	1.293	0.161	8.050	<.01
	grade 3	0.519	0.178	2.915	<.01		useepst	-1.335	0.132	-10.103	<.01
	useepst:grade 2	0.384	0.226	1.698	0.090		grade 2	0.421	0.155	2.727	<.01
	useepst:grade 3	0.521	0.228	2.283	<.05		grade 3	0.595	0.156	3.817	<.01

4. 結果と考察

4.1 RQ1 一般化線形混合モデルと推定周辺平均値の有用性について

3.3 節に従いモデル推定を行ったところ, “can”と“will”, “would”では, 交互作用を考慮した model 2b が最適なモデルであると推定された。他の法助動詞については, 交互作用を考慮しない model 2a が最適なモデルであると推定された。その後, 全ての法助動詞の推定周辺平均値を算出し, 各意味別頻度と学年別の頻度の間で有意差があるかを調べた。

交互作用を考慮したモデル結果を, 図 1 から図 3 に示す。“can”の dynamic な意味や epistemic な意味のように, 各法助動詞の各意味を学年別で比較すると, 1 年生と 2 年生, 3 年生の間で有意差が見られた。学年が上がるにつれて, これらの法助動詞の頻度が高くなることを示している。また, 各学年の頻度を意味別に頻度を比較すると, 全ての意味の間にも有意差があった。学年を通して, 頻繁に使われる意味とそうでない意味が明確であることが明らかになった。

一方で, 交互作用を考慮しない法助動詞を分析した結果, 各法助動詞の, 全ての意味の組み合わせにおいて p 値が.01 以下であり, 有意差があった。また, 学年別で比較した結果, 表 4 のとおり, 1 年生と 2・3 年生の間で有意差が多く確認された。このことから, 法助動詞は 1 年生では相対的に扱う頻度が少なく, 2 年生と 3 年生でより集中的に学習されると考えられる。

図 1 can の推定周辺平均値の多重比較結果

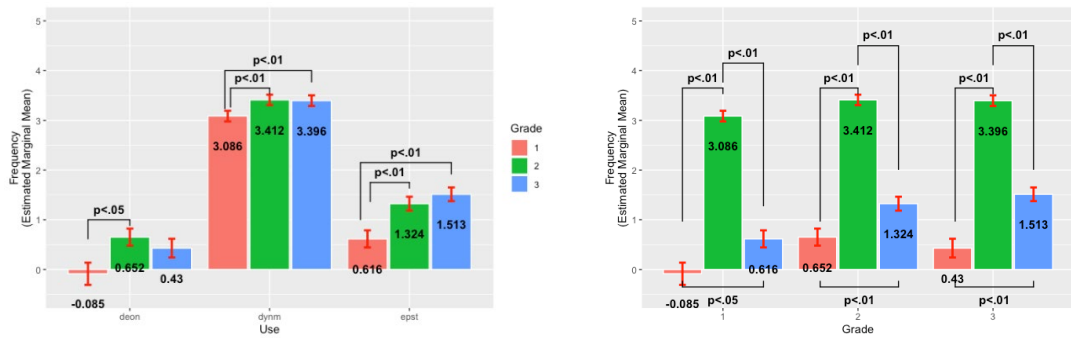


図 2 will の推定周辺平均値の多重比較結果

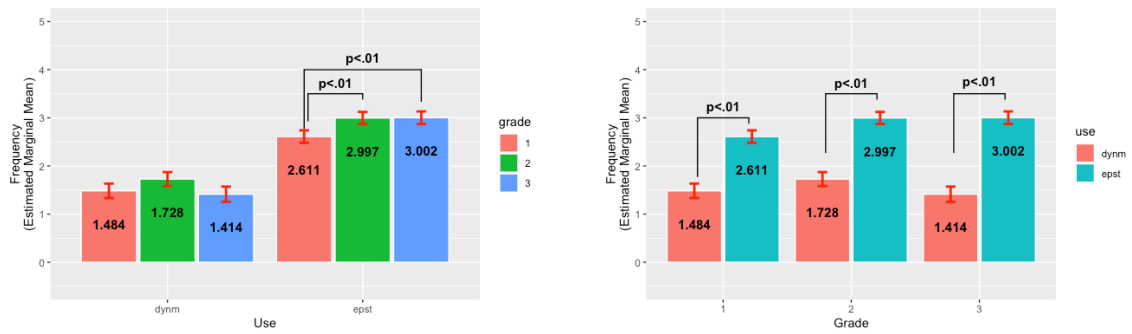


図 3 would の推定周辺平均値の多重比較結果

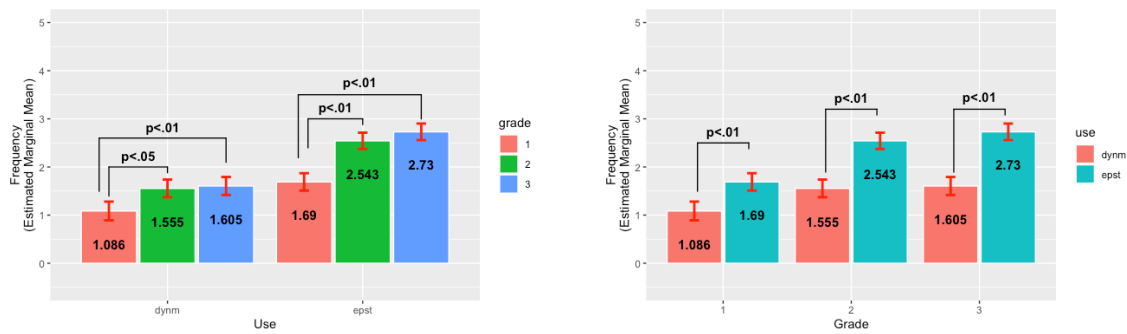


表 4 could, may, might, must, should の学年別の結果

法助動詞	学年	p値	法助動詞	学年	p値
could	1 – 2	<.01	must	1 – 2	0.399
	1 – 3	<.01		1 – 3	<.05
	2 – 3	0.08		2 – 3	0.363
may	1 – 2	0.131	should	1 – 2	<.05
	1 – 3	<.01		1 – 3	<.01
	2 – 3	<.01		2 – 3	0.472
might	1 – 2	<.01			
	1 – 3	<.01			
	2 – 3	0.131			

このように、一般化線形混合モデルと推定周辺平均値を用いた分析では、交互作用を考慮すべき法助動詞とそうでないものを確認できた。また、交互作用を考慮すべき用例では、学年が上がるにつれて頻度も増えることが確認でき、高校教科書の特徴を捉えることができた。さらに、交互作用を考慮しない場合も、意味別と学年別の両方で有意差を確認することができた。表 1 のデータで、意味別にも学年別にも頻度が偏っていることから、妥当な結果が得られたと考えられる。

5. まとめ

本研究では、小規模コーパス分析において、一般化線形混合モデルと推定周辺平均値を用いた分析の有用性を考察した。まず、法助動詞を分析する際に、交互作用を考慮すべきかどうかを判断することができた。そして、考慮すべき場合において、学年が上がるにつれて意味別頻度が高くなることがわかり、交互作用を確認することができた。また、交互作用を考慮しない場合においても、粗頻度データから予想される結果を確認することができた。これらの結果から、小規模コーパス分析に対して一般化線形混合モデルと推定周辺平均値は有用であると考えられる。

引用文献

- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. 10.1017/9781316410899
- Imao, Y. (2021). CasualConc (Version 2.1.7). <https://sites.google.com/site/casualconc/>
- Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, 1, 91–118.
- 梶山達也 (2022.3) 「教科書コーパスを用いた英語助動詞の頻度と意味についての探索的研究」『英語コーパス学会語彙研究会』, <http://jaecs-lexsig.blogspot.com/>
- 水本篤 (2010) 「サンプルサイズが小さい場合の統計的検定の比較: コーパス言語学・外国語教育学への適用」『言語コーパス分析における数理データの統計的処理手法の検討』, 1–14.
- 森秀昭 (2020) 「学習者コーパスに適した分析法の検討」『計量国語学』32(7), 436–446.
- Newbery-Payton, L. (2022). A learner corpus-based study of L1 effects on L2 English auxiliary verb use: The case of *will*. *English Corpus Study*, 29, 19–40.
- Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, 34(4), 216–221. 10.1080/00031305.1980.10483031

分布意味論の手法を応用したフレーム意味論の分析

森下 裕三(桃山学院大学)
ymorishi@andrew.ac.jp

A Frame Semantics Analysis with the Use of Distributional Semantic Methods

MORISHITA Yuzo (Momoyama Gakuin University)

Abstract

The purpose of this study is to make use of the methods of distributional semantics (e.g., Firth 1957, cf. Harris 1954) to verify what have been argued within the theoretical framework of Frame Semantics (e.g., Fillmore 1982). Using randomly extracted examples from the Corpus of Contemporary American English (COCA), the semantic relations of four nouns (land, sea, ground, and air) and four adjectives (stingy, thrifty, wasteful and generous) were clarified. The results generally support what have been argued in Frame Semantics, although it is difficult to properly analyze homonyms and polysemous words in type-based semantic vector spaces.

Keywords

フレーム意味論, 分布意味論, 形容詞, 名詞

1. はじめに

認知言語学の特徴のひとつとして、意味に認識主体の存在が関与することが挙げられる。この意味観は、Fillmore を代表とする研究者らがフレーム意味論として発展させてきた (e.g., Fillmore 1982)。語の背景にある豊かな知識を意味分析に利用できるようになり、land と ground といった語の違いまで分析できるようになった (Fillmore 1982: 120-121)。

フレームという膨大な知識をコーパス研究と結びつけるプロジェクトも存在するが (e.g. Baker, Fillmore and Lowe 1998), 本研究では分布意味論と呼ばれる探索的な手法によって、フレーム意味論の枠組みから議論されてきた語が分析可能であることを示す。現代のアメリカ英語を代表する Corpus of Contemporary American English (COCA) のテキストデータを利用して、先ほどあげた land と ground の他に stingy, thrifty, wasteful, generous という語の意味分析をおこなう。

本稿の構成は次のとおりである、2 節ではフレーム意味論や分布意味論といった本研究の理論的な背景について概観する。さらに、3 節では本研究における分析手法などについて説明し、4 節では分析結果を示す。さらに、5 節では本研究のまとめと課題を述べる。

2. 先行研究

ここからは、本研究の理論的背景である 2 種類の意味論を紹介しながら、これらの意味論の接点を模索する。

2.1 フレーム意味論

1 節でも触れたとおり、フレーム意味論は 1980 年代から Fillmore を中心とした研究者らが発展させてきた理論である。語の意味分析や意味記述のために、認識主体が持つ語の背景知識が必要不可欠であるとした理論である。先にみた **land** と **ground** であれば、これらの語は客観的には地表という同じモノを指すことになる。しかし、**land** が認識主体に **sea** との対比を求めるのに対して、**ground** は認識主体に **air** との対比を求める。この違いは、以下の各文で示されるのがどのような鳥なのかといった違いにもあらわれている。

(1) a. a bird that spends its life on the land

b. a bird that spends its life on the ground (Fillmore 1982: 121)

他の例として形容詞の **stingy**, **thrifty**, **wasteful**, **generous** という語がある。あまりお金を使わないという行為や態度に対して否定的な評価をするのが **stingy**, 肯定的な評価をするのが **thrifty** である。そして、お金をよく使うという行為や態度に対して否定的な評価をするのが **wasteful**, 肯定的な評価をするのが **generous** である。

2.2 分布意味論

分布意味論というのは、周辺に共起する語から任意の語の意味が推測できるとする理論である。コーパスの充実や計算機の高まるなかで、計算言語学や自然言語処理など言語学と隣接する分野でも拠り所とされている (e.g., Milkolov et al. 2013)。

多くの研究で応用されてはいるが、利用するコーパスや設定するパラメータによって分析結果が変化するという課題もある。実際に、任意の語と共起する語だけでなく、任意の語と共起する語に共起する語という非常に膨大なデータを分析対象とするといった試みもみられる (e.g., Hilpert and Saavedra 2020)。

本研究では次節で紹介するような方法に基づいて、認識主体によって喚起される多くの知識が分布意味論に基づく分析手法によって分析できる可能性を示す。

3. データ

本節では、前節までで説明してきた考え方に基づいて、COCA のデータを利用して、まず、**land** と **ground** の意味分析をおこなう。次に **stingy**, **thrifty**, **wasteful**, **generous** の意味分析を

おこなう。

最初に意味分析をおこなう **land** と **ground** は、フレーム意味論の考え方が分布意味論による分析手法にも反映されていれば、**land** は **sea** と近似し、**ground** は **air** と近似する結果が得られるはずである。そして、**stingy**, **thrifty**, **wasteful**, **generous** は、行為や態度についての否定的な評価が分布意味論による分析手法にも反映されていれば、**stingy** と **wasteful** が近似し、行為や態度についての肯定的な評価が反映されれば **thrifty** と **generous** が近似するという結果が得られるはずである。

3.1 データ分析のためのパラメータと意味ベクトル

本研究では、まず i) COCA のテキストデータに生起した **land**, **sea**, **ground**, **air** という語の全用例からランダムにそれぞれ 100 例を抽出した。次に、ii) ランダムに抽出された各語の用例の前後 5 語ずつのみを取り出した。さらに、iii) 語の意味にあまり影響がないと考えられる冠詞、数字、接続詞をすべて取り除いた。このような手続きの結果、以下の表 1 のように、それぞれの語に共起する語の頻度をベクトル化した。

図 1: 名詞の意味ベクトル

	abalone	[...]	by	[...]	off	[...]	on	[...]
land	0		6		0		10	
sea	0		5		2		17	
ground	1		6		1		13	
air	0		2		2		40	

3.2 自己相互情報量

前節でみた名詞の意味ベクトルは頻度情報のみが記載されたものであるため、この情報だけで各語の意味的な近接性を計算することはできない。そのため、i) まず、各語の共起頻度の情報から、共起頻度の期待値を計算し、ii) 自己相互情報量 (Pointwise Mutual Information Score) を計算し、iii) 各語の類似度を計算した (Levishina 2014: Chapter 16)。

このような手順によって、**land**, **sea**, **ground**, **air** という名詞の意味と、**stingy**, **thrifty**, **wasteful**, **generous** という形容詞の意味を分析し類似度を計算した。

4. 結果と考察

4.1 名詞の意味分析についての結果と考察

名詞である **land**, **sea**, **ground**, **air** の分析結果では、フレーム意味論が想定していた通り **land** と **sea**, そして **ground** と **air** に高い類似度がみられた。このことから、ある程度の量のデータがあれば、フレーム意味論に基づく語と語の関係も分布意味論の手法によって分析できる可能性を示

すことができた。しかし、実際には問題もあり、たとえば、**ground** という語は、数例のみではあるが、動詞 **grind** の過去分詞形が含まれており、品詞情報を取り除いた分析の限界も示すことになった。

4.2 形容詞の意味分析についての結果と考察

先にみた **land, sea, ground, air** とは異なり、形容詞の分析ではより大きな問題が見つかった。たとえば、**generous** という語は金銭の支払いに対する肯定的な評価をあらわす意味以外にも多くの意味がある。そのため、ランダムに抽出された例のうち多くの例が、意図していた意味をあらわすものではなく、妥当な分析結果が得られなかった。

5. まとめ

本研究では、一部で問題は見られたものの、フレーム意味論が想定していた結果を得ることができた。しかし、語の多義性が関与する場合は分析に失敗するという可能性も示唆することができた。今後は、この課題を克服すべく各トークンの分析が必要である。

しかし、タイプの分析からも多くの情報が得られるため、同音意義語や多義語が関与しない領域の類義語や反義語などを分析する場合はタイプによる分析も有用である可能性が高い。また、タイプの分析で得られた結果とトークンの分析で得られた結果を比較することで分析の問題点を見つけ出すこともできるだろう。

引用文献

- Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The Berkeley FrameNet project. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1*, 86-90.
- Davies, M. (2010-) The Corpus of Contemporary American English (COCA)
- Fillmore, C. J. (1982). Frame semantics. In The Linguistic Society of Korea (ed.), *Linguistics in Morning Calm*, 111-137. Seoul: Hanshin Publishing Company.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. Philological Society (ed.) *Studies in Linguistic Analysis*, 1-32.
- Harris, Z. (1954). *Distributional Structure*. In Word, 10, 146-162.
- Hilpert, M. and D. C. Saavedra (2020). Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*.1-32
- Levshina, N. (2014) *How to do Linguistics with R data exploration and statistical analysis*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Mikolov, T. K. Chen, G. Corrado and J. Dean (2013) Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

Negated speech and thought presentation in contemporary present-tense fiction

IKEO Reiko (Senshu University)

rikeo0919@gmail.com

Abstract

This paper shows how negated speech and thought presentation functions in narrative by using a corpus approach. A corpus consisting of texts from contemporary present-tense fiction was annotated with discourse presentation categories based on the Semino and Short model (2004). In addition, the new subcategory ‘g’ has been introduced for annotating negated discourse presentation. The data shows that negated thought presentation occurs more frequently than negated speech presentation. Negated cases of speech and thought presentation reveal characters’ inner worlds in a way that affirmatives do not.

Keywords

corpus stylistics, negated speech presentation, negation, present-tense fiction, thought presentation

1. Negated speech and thought presentation

In fictional texts, a character’s speech can occasionally appear in the form of negation. In speech presentation, negation often connotes that the speech event referred to had been expected or assumed but did not actually happen as shown in the example:

- (1) Despite the shared and joshing friendship of our youths, *I’ve never said, Your roof can wait. There’re our own oxen need attending to.*
(Jim Crace, *Harvest*, 2013:54-55, my italicisation)

In narrative, character speech, thought and writing are normally presented in affirmatives as new information and contribute to developing the story. Givón (1978) states that negatives involve more presuppositional complexity and a greater abundance of assumptions which the speaker must make about what the hearer knows than their corresponding affirmatives. Negatives are used in a context where corresponding affirmatives have already been discussed, or the speaker assumes the hearer is familiar with the corresponding affirmative. Hidalgo Downing (1991: 147) explains that negation

is a natural foregrounding device in narrative texts. When non-events or non-states are mentioned because they are considered to be more informative than events or states, Hidalgo Downing argues, they stand out and have special textual effects. This paper shows how negated speech and thought presentations function in contemporary literary texts by using a corpus approach.

2. The corpus and its annotation

The corpus for this research consists of texts from contemporary present-tense fiction. It comprises a collection of forty extracts from fiction published between 2000 and 2016, each consisting of approximately 2,000 words. The total word count of the corpus is 89,912. The corpus was annotated with speech, writing and thought presentation categories developed by Semino and Short (2004).

Table 1 Categories and examples of the speech presentation cline

Categories	Examples
FDS (Free Direct Speech)	‘The flat of the three blind mice?’ (M. J. Hyland, <i>Carry Me Down</i> , 2006: 247)
DS (Direct Speech)	‘But why me, Cornelia?’ she asks. (Jessie Burton, <i>The Miniaturist</i> , 2014: 235)
FIS (Free Indirect Speech)	It has a stain on the hem but she says it would suit me. <i>She’ll mend it when she gets round to it because she knows my medusa stings hurt all the time.</i> (Deborah Levy, <i>Hot Milk</i> , 2016: 101)
IS (Indirect Speech)	Futh mentions to Carl that he has not been driving very long. (Alison Moore, <i>The Lighthouse</i> , 2012: 26)
NRSA (Narrator’s Representation of Speech Acts)	Anne the queen sends for him to her own chamber. (Hilary Mantel, <i>Bring Up the Bodies</i> , 2010: 74)
NV (Narrator’s Representation of Voice)	She waves, <i>says a few more words to the man</i> and comes over to me. (Clare Morrall, <i>Astonishing Splashes of Colour</i> , 2003: 168)

The Semino and Short discourse presentation model has three separate clines for speech, writing and thought presentation. Each cline has six major categories and a separate category for reporting clauses. The categories of the speech presentation cline

and examples from the corpus are shown in Table 1. As the list descends, the narrator's control over the character's discourse increases. The first two categories FDS and DS are direct forms, while the rest are indirect forms. The thought presentation cline has six major categories similar to those of the speech presentation cline, such as Free Direct Thought (FDT) and Indirect Thought (IT). All the texts in the corpus are annotated with these main categories, and segments which do not belong to any discourse presentation are annotated as "narration." For negated speech and thought presentation, the subcategory "g" is attached to the main category.

3. Findings

In speech presentation, the negated speech events referred to were either assumed to be happening or had been expected but did not actually happen in the narrated world. In Example (1) previously mentioned, FDS is negated and in (2) below IS is negated:

(2) ISg

Hmm. Wait a sec. *I didn't actually ask her if I could tell you this*, and now that I think about it, maybe I should ask her first.

(Ruth Ozeki, *A Tale for the Time Being*, 2013: 178, my italicisation)

Table 2 *Raw frequency of "g" in speech presentation*

Main categories	total no. of tags	Raw frq of "g"
FDS	1426	2
DS	633	0
FIS	14	0
IS	97	4
NRSA	228	11
NVg	111	6
NRS (reporting clauses)	752	6
Total	3261	29

Syntactically, direct forms of speech presentation cannot be negated unless a reported clause is accompanied by a reporting clause such as "he says." This can also be applied to cases of direct thought presentation. The corpus data suggests that speech presentation is very rarely negated both in direct forms and indirect forms in fictional texts. Table 1 shows the raw frequency of "g" attached to the main categories on the speech presentation cline. The subcategory "g" is attached to 29 out of 3,261 speech

presentation cases, which account for only 0.9% of all the speech presentation tags. Out of 29 cases, there are only two direct forms that are negated, and the rest are indirect forms. Pragmatically, negated speech presentation invites the reader's inferences about the non-events mentioned in the texts: in Example (1) why the character felt like speaking out but did not do so and in Example (2) why she mentions the non-event in the past and her intention to achieve it now. By processing this information, the reader can infer what the relationships between the speakers and addressees are and what personalities the speakers possess.

Negated thought presentation occurs more frequently compared with speech presentation. As Table 2 shows, the "g" subcategory is found in 150 cases out of 1,805 thought presentation cases, which account for 8.3% of the occurrences of thought presentation.

Table 3 *Raw frequency of "g" in thought presentation*

Main categories	total no. of tags	Raw frq of "g"
FDT	449	0
FIT	123	0
IT	261	42
NRTA	223	37
NI (Internal states)	428	30
NRT (Reporting clauses)	321	41
Total	1805	150

No direct forms of negated thought presentation are found, and all the negated cases are indirect forms. IT has the largest number of cases, followed by NRTA and NI. In thought presentation, the scope of negation seems to be dependent on cognitive verbs which introduce a character's thoughts. Cognitive verbs are divided into two types: factive verbs and non-factive verbs. Factive verbs such as "know," "remember," "understand," and "forget," denote a true event or a fact, which is described in the complement clause. When a factive verb is used for thought presentation, the proposition of the reported clause is shielded from the negation:

(3) ITg (Indirect Thought)

Sam can't remember what her mother had been so furious about,
(Miranda Shelly, *Bone Meal for Roses*, 2017: 225)

On the other hand, when character thought is presented through a non-factive

cognitive verb such as “think” and “believe,” which does not presuppose the truth of the proposition expressed in the that-clause which follows, the content of the that-clause is affected by the negation.

(4) ITg

I don't think I'll be going past town anymore.

(DBC Pierre, *Vernon God Little*, 2003: 137)

When expressions which are frequently used in negated thought presentation are divided in terms of factive and non-factive traits, approximately 30% of the cases are factive, including “know” “remember” and “understand” whereas approximately 70% are non-factive, including “think” “want” and “be sure”.

In many cases of indirect thought presentation, the negated segments deny a certain psychological state and imply another, often a state of lacking cognition or intention, or an unusual state of mind affected by external stimuli, such as being upset, confused, sad or surprised rather than presupposing an affirmative as in negated speech presentation. The following examples have negations in NRTA and NI respectively:

(5) NRTAg (Narrator's Representation of Thought Acts)

I have no idea what to do with him

(Cecelia Ahern, *Thanks for the Memories*, 2008: 234)

(6) NIg (Internal Narration)

But suddenly, I'm not so sure.

(Giselle Green, *Dear Dad*, 2016: 279)

Negated speech and thought presentation can reveal important linguistic and pragmatic differences between speech presentation and thought presentation. Negated thought presentation occurs eight times more often than negated speech presentation, and they are more often found in 1st-person narration than in 3rd-person narration. The most frequent cognitive verbs used in negated thought presentation are “know” (16 times) followed by “think” and “want” (11 times each). These verbs are used in multi-word clusters most frequently found in spoken discourse such as “I don't know (what/how/if/whether),” “I don't think,” or “I don't want (Carter & McCarthy 2006).” This suggests that characters' mental states are often presented by the 1st-person narrator using these phrases, which are commonly found in spoken discourse. In 3rd-person

narration similar expressions are found in many cases with the syntactic subject being substituted for 3rd-person pronouns.

Negated speech presentation seems to be much less common in fictional texts. This suggests that speech is usually presented as an event rather than a non-event. However, when negated speech presentation occurs in narrative, which is linguistically less common than in thought presentation, it is deviating from the norm, and therefore, can impose upon the reader the need to expend more effort to interpret the negated speech presentation than is required for negated thought presentation.

4. Conclusion

This paper has examined negated speech and thought presentation based on corpus data. The frequency of negated thought presentation is greater than that of negated speech presentation. This suggests that describing a character's internal states in terms of negation is a practice more commonly adopted in narrative. Many of the expressions are often found in non-fictional spoken discourse. The combination of a negated reporting clause and an indirect form of thought presentation suggests a character's confused, upset or disoriented internal state at a specific moment in time, rather than revealing the character's presuppositions or assumptions. On the other hand, a speech event presented as a non-event seems to be less frequent in narrative. When negated speech presentation is used, it reveals the presuppositions and assumptions which are involved in the presented speech. Such presuppositions and assumptions can be exploited to reveal characters' inner worlds which cannot be divulged by descriptions of actions or events.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 21K00573.

Bibliography

- Carter, R. & McCarthy, M. (2006). *Cambridge grammar of English*. Cambridge University Press.
- Givón, T. (1978). Negation in language: pragmatics, function, ontology. In P. Cole (Ed.), *Syntax and semantics 9: pragmatics*, (pp.69-112). Academic Press.
- Hidalgo Dowling, L. (2000). *Negation, text worlds, and discourse: the pragmatics of fiction*. Ablex Publishing Corporation.
- Semino, E. and Short, M. (2004). *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing*. Routledge.

『英語コーパス学会大会予稿集 2022』（ISSN 2436-6447）

Proceedings of the JAECS Conference 2022

刊行日 2022 年 10 月 1 日

発行所 英語コーパス学会

事務局 〒501-0192 岐阜市一日市場北町7番1号岐阜市立女子短期大学英語英文学科 小島ますみ研究室気付

e-mail: jaecs.hq@gmail.com URL: <http://jaecs.com/>