英語コーパス学会
大会予稿集2021

PROCEEDINGS OF THE JAECS CONFERENCE 2021

2021.10.2

ISSN 2436-6447

JAECS
Japan Association for English Corpus Studies

目次

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

知覚動詞補文に出現する受身表現の容認可否について

村岡　宗一郎（日本大学 大学院生）
hollow_t_classic@ezweb.ne.jp

On the Acceptability of Passive Expressions in the Complement of Perception Verbs

MURAOKA　Souichiro (Nihon University, Graduate Student)

Abstract

Regarding the use of the non-finite verbs in the complement of perception verbs, “be + -en” is unacceptable, while “being + -en” and “get + -en” are acceptable. However, unacceptable forms such as “see NP be + -en” are used in practice, for example, “I couldn’t stand to ***see her be cremated***.(Murakami Haruki, Killing Commendatore.)”. This study analyzes how often these example such as “see NP be + -en” are used in reality and what semantic constraints are imposed upon them by examining data from corpora, such as BNC and COCA. This study confirms that “see NP be + -en” is used more often in American than in British English, and “see NP get + -en”, which has been considered by previous studies to be grammatically correct, is also mainly used in American English. I conclude that the use of “see NP be + -en” has increased along with the use of “see NP get + -en” in American English.

Keywords

知覚動詞, 補文, 準動詞, 受身表現

1. はじめに

知覚動詞補文は（1）に示すように，原形不定詞，現在分詞，過去分詞を補文にとる。このうち，原形不定詞は知覚事象の完結性を，現在分詞は知覚事象の非完結性および一時性を表すとされている（cf. Allen 1974[4]: 186）。

（1） a. I ***saw the children eat*** their lunch.　（Palmer 1987[2]: 199）
　　b. I ***saw the children eating*** their lunch.　（ibid.）
　　c. I ***saw the children beaten*** by their rivals.　（ibid.）

このうち，（1c）は受身を表すが，類似する表現として，（2）のような表現も存在する。

(2) a. *I ***saw him be rejected***. (Bolinger 1974: 69)
b. I ***saw the children being beaten*** by their rivals. (Palmer 1987[2]: 199)
c. I ***saw him get rejected***. (Bolinger 1974: 69)

しかし，(2a)の原形不定詞補文における受身表現は，多くの先行研究で非文法的とみなされている。また(2a)が非文法的とみなされるのに対して，(2c)は容認されている。これは，状態受身と動作受身の違いに起因するものと考えられる。前述の通り，原形不定詞は知覚事象の完結性を表すが，吉良(2006: 46)によれば，(3)のような知覚動詞補文における状態動詞の出現は，「状態的な出来事」は終結点を持つ「完了した事象」とは捉えられず，容認できないと述べる。

(3) a. *We ***saw John look*** pretty sick. (Akmajian 1977: 440)
b. *I ***saw Tom still resemble*** your father. (Declerck 1981: 89)

実際に(4)のような例が確認されるが，安藤(2005: 829)によれば，稀な例であるという。しかし，安藤の見解の拠り所は，Palmer(1968: 171)にあるとされているが(cf. 安藤(2005: 829-830))，Palmer(1974)や(1987[2])からはその記述は削除されている。

(4)I couldn't stand to ***see her be cremated***. (Murakami Haruki, Killing Commendatore.)

本研究では，(4)のような知覚動詞の原形不定詞補文における受身表現はどれほど容認されているのか，またどのような意味的制約が課されているのか，BNC や COCA を用いて分析を行う。

2. 先行研究

(2a)の容認可否について，(2a)を文法的と見なす先行研究は少数派であり，大多数の先行研究において，非文法的とみなされている(cf. Bolinger(1974: 69)，Lapointe(1980: 772)，Declerck(1991: 490)，Clark and Jäger (2000: 19))。(2a)が非文法的とみなされる要因については，前述の通り，be が状態受身として解釈されやすいためであると述べたが，Bolinger (1974)は(5)のように習慣や反復を表す場合には，be + en 補文も容認されると述べる。これは個々の状態的な出来事が，繰り返しの動作として捉えられるためである。

(5) a. I <u>used to</u> ***see him be rejected***. (Bolinger 1974: 69)
b. <u>Again and again</u> I ***saw him be rejected***. (ibid.)

また知覚動詞においては，Martha ***saw the policeman be mammals***. の様な個別レベル述語を用いた例はその事象を知覚できないため容認されないが(cf. Carlson(1977: 125))，白井(1999: 20)によれば，(6a)のように振る舞うというような動作的な意味の be であれば容認可能で

あるという。また Felser（1999: 83）は（6b）のように場所を表す副詞句と共起し，状態性を弱めることで，原形不定詞補文における be + -en の出現は容認され，中右（1980: 147）も同様に，非状態的な事態を叙述している場合には（6c）の例は容認されうるという。

（6） a. We ***saw John be polite*** for the first time. （白井 1999: 20）
b. We ***saw John be drawn*** into the game. （Felser 1999: 83）
c. I don't like to ***see people*** {***be*** / ***being***} ***intimidated***. （中右 1990: 147）

また Bolinger（1974）や柏野（1993）は，知覚動詞の完了形であれば，原形不定詞補文においても，be + en は容認されるという。

（7） a. I ***have seen him be rejected***. （Bolinger 1974: 69）
b. I ***'ve*** never ***seen a man be executed*** before. （柏野 1993: 81）

（7）の現在完了には「完了・結果」と「経験」の二つの解釈が可能であるが，柏野（1993: 78）によれば，主文が完了形で「完了・結果」の意味の場合には，主文に示される知覚過程の完結を強調するので，補文の行為も終わっていることになり，原形不定詞が選ばれるという。その一方で，補文の動詞が状態動詞で主文の動詞の時制が「経験」を表す完了形であれば，状態動詞（be を含む）も補文に生起可能となるという（cf. 柏野（1993: 79, 81））。しかし，Gee（1975）は，（8）の例を挙げ，（7）の例が「完了」の読みになり得る可能性を否定している。Gee（1975: 377）によれば，「ある特定の場面で実際に（目の前で）知覚した出来事」を表す場合には，I ***saw a car be wrecked*** by the police. のような例は容認されないという。そして，（8a）の現在完了形を「完了・結果」用法と解釈できる場合には，過去形の saw とほぼ同じであり，「ある特定の場面で実際に（目の前で）知覚した出来事」を表すため容認されないが，（8b）のように補文主語を複数形にして，現在完了形を「経験」用法と解釈できる場合には「異なる場面で繰り返された出来事」を表すため，容認されるという。

（8） a. I ***'ve seen a car*** (#***be***) ***wrecked*** by the police. （Gee 1975: 377）
b. I ***'ve seen cars be wrecked*** by the police. （ibid.）

以上の先行研究をまとめると，知覚動詞補文における be + -en の出現は，be が動作的なものや知覚動詞の完了形が経験を表す場合には容認される。この点について，コーパスを用いて実証していく。

3. リサーチデザイン

前節において，（2a）に見られる表現が容認されうる環境について，先行研究の見解を見てきたが，動作性を表す be + -en や経験を表す知覚動詞の完了形のような例は実際にどれほど存在

するのかについて，BNC や COCA を用いて調査を行う。検出方法としては，知覚動詞の補文主語のバリエーションを考慮し，(9)に示す 13 パターンの検索式を用いて調査を行った。また being + -en，先行研究で容認されている get + -en，getting + -en や過去分詞補文も同様に，その割合を調査した。

(9) a. {[see] / [hear] / [watch]} (ART/DET) NOUN be _v?n
b. {[see] / [hear] / [watch]} PRON be _v?n
c. {[see] / [hear] / [watch]} (ART/DET) ADJ NOUN be _v?n
d. {[see] / [hear] / [watch]} (ART/DET) NOUN NOUN be _v?n
e. {[see] / [hear] / [watch]} (ART/DET) ADJ NOUN NOUN be _v?n

4. 結果と考察

前節で見た調査方法をもとに調査を行った結果，表 1 の結果が得られた。全体の割合としては，英米共に be + -en 補文は少数ではあるが，アメリカ英語に多く確認された。特にアメリカ英語の watch において，be + -en 補文と get + -en 補文がその他知覚動詞よりも高い割合で検出された。これは，watch が動作性のある知覚対象を目的語にとるためであると考えられる。

表 1. BNC および COCA の知覚動詞補文における受身表現とその分布

		BNC		COCA	
see	**be -en**	**4**	**0.3%**	**118**	**0.8%**
	being -en	233	17.5%	1698	11.5%
	get -en	**3**	**0.2%**	**647**	**4.4%**
	getting -en	3	0.2%	174	1.2%
	-en	1092	81.8%	12131	82.1%
hear	**be -en**	**0**	**0%**	**18**	**0.9%**
	being -en	50	23.5%	187	9.8%
	get -en	**0**	**0%**	**12**	**0.6%**
	getting -en	1	0.5%	14	0.7%
	-en	162	76%	1683	87.9%
watch	**be -en**	**0**	**0%**	**199**	**10.7%**
	being -en	71	51.1%	509	27.3%
	get -en	**1**	**0.7%**	**220**	**11.8%**
	getting -en	1	0.7%	19	1%
	-en	66	47.5%	918	49.2%

また先行研究では知覚動詞の完了形に後続する補文において，be + -en が用いられると述べられていたが，そのような用例は多くは検出されず，(10)に示すような (don't) want to や be {glad/sad} to などの意志や感情を表す語句との共起がより多く確認された。

(10) a. I don't want to ***see any woman be misdiagnosed*** and have to live with what I've been told

and what I know right now. （COCA, 1990. SPOK）

b. I'd hate to ***see them be slimmed*** down into fewer. （COCA, 2012. BLOG）

c. I'd like to ***see writers get paid*** fairly and the library model in some other countries actually does compensate them for when their books are lent in libraries （COCA, 2012. BLOG）

d. I hate to ***see people get sucked*** into the idea that they're going out serving the Lord, （COCA, 1993. NEWS）

さらに，先行研究で文法的と見なされていた get + -en 補文に関しては，ほぼアメリカ英語にのみ確認される。元々，get 受動文はイギリス英語よりもアメリカ英語に多く見られたとされており（cf. Sussex（1982: 90)），また 20 世紀には主として米語の口語的表現において大量に用いられると（cf. 松元（2011: 22)），20 世紀後半には書き言葉にも多く用いられるようになったとされている（cf. Schwarz（2017; 2019)）。これらの調査結果から，be + -en 補文の使用における英米の差は get 受動文および get + -en 補文の定着が大きく影響を及ぼしている可能性が考えられる。Felser（1999: 82)もまたイギリス人英語母語話者は，be + -en 補文の使用を避ける傾向にあるとする一方で，アメリカ人英語母語話者の多くは違和感を覚えることもあれば，文法的とみなす者もいるというが，アメリカ英語における get 受動文の確立に伴い，知覚動詞補文にも同様に，get + -en 補文が用いられるようになると，それまで非文法的とみなされていた be + -en 補文もまた類似表現として徐々に用いられるようになったと考えられる。

5. まとめ

先行研究の多くでは，動作受身を表す場合や知覚動詞が現在完了形で用いられている場合を除いて，知覚動詞の be + -en 補文は一般的に非文法的と見なされていた。このことについて，BNC や COCA を用いて調査を行った結果，知覚動詞の be + -en 補文は，全体の割合としては少数ではあるが，主にアメリカ英語に多く確認された。また先行研究では容認されていた get + -en 補文もまたアメリカ英語にのみ検出された。元来，get 受動文はアメリカ英語の口語表現であったものが，時代を経るにつれて確立していったとされるが，アメリカ英語の知覚動詞補文も同様に，get + -en 補文の使用が定着すると，それまで非文法的と見なされていた be + -en 補文もまた類似する表現として徐々に用いられるようになったと考えられる。

引用文献

Akmajian, A. (1977). The complement Structuture of Perception Verbs in an Autonomous Syntax Framework. In Culicover, A, W., Akmajian, A. and Wasaw, T. (Eds.), *Formal Syntax*. Academic Press, 427-60.

Allen, W. S. (1974[4]). *Living English Structure*. Longman.

安藤貞雄 (2005)『現代英文法講義』開拓社.

Bolinger, D. (1974). Concept and percept: Two infinitive constructions and their

vicissitude. *World Papers in Phonetics Festschrift for Dr. Onishi's Kiju*. (pp. 65-91). The Phonetic Society of Japan.

Carlson, G. N. (1977). *Reference to Kinds in English*. Doctoral Dissertation. MIT.

Clark, R. and G. Jäger. (2000). A Categorial Syntax for Verbs of Perception. University of Pennsylvania. *Working Papers in Linguistics* 6: Iss. 3, Article 5, 15-33.

Declerck, R. (1981). On the role of progressive aspect in nonfinite perception verb complements. *Glossa* 15, 83-114.

Declerck, R. (1991). *A comprehensive descriptive grammar of English*. Kaitakusha.

Felser, C. (1999). *Verbal Complement Clauses: A Minimalist Study of Direct Perception Construction*. John Benjamins Publishing Company.

Gee, J. P. (1975). *Perception, Intentionality, and Naked Infinitives: a Study in Linguistics and Philosophy*. Doctoral dissertation, Stanford University.

柏野健次 (1993)『意味論から見た語法』研究社.

Lapointe, S. G. (1980). A note on Akmajian, Steel and Wasow's treatment of certain verb complement types. *Linguistic Inquiry* 11, 770-787.

松元浩一 (2011)「18 世紀英語の get-受動文」『長崎大学教育学部紀要:人文科学』(77), 21-35.

中右実 (1980)「テンス, アスペクトの比較」國廣哲彌 (編)『日英語比較講座第 2 巻 文法』大修館書店.

Palmer, F. R. (1965). A Linguistic Study of English Verb. Longman.

Palmer, F. R. (1974). *English Verb*. Longman.

Palmer, F. R. (1987[2]). *English Verb*. Longman.

Schwarz, S. (2017). Like Getting Nibbled to Death by a Duck: Grammaticalization of the Get-passive in the TIME Magazine Corpus. *English Word-Wide*, 37(3), 305-335.

Schwarz, S. (2019). Signs of Grammaticalization. Tracking the Get-passive through COHA. In Claridge, S. and Bös, B. (Eds.), *Developments in English Historical Morpho-Syntax*. 199-222. John Benjamins.

白井賢一郎 (1999)「英語の知覚動詞構文：視覚情報モデルに基づく認知的研究」『中京大学教養論叢』(40), 64-67.

Sussex, R. (1982). A note on the get-passive construction. *Australian Journal of Linguistics* 2, 83–95.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 2021

「1961-2021 日本語小説コーパス」の構築
―日英小説対照研究の新しい可能性―

石川 慎一郎(神戸大学)
iskwshin@gmail.com

"1961-2021 Japanese General Fiction Corpus"
―For a New Comparative Study of Japanese/ English Fictions―

ISHIKAWA Shin'ichiro (Kobe University)

Abstract

This paper illustrates the design of the "1961-2021 Japanese General Fiction Corpus," which is a collection of Japanese fiction works published in 1961, 1971, 1981, 1991, 2001, 2011, and 2021. It also includes two kinds of English translations. This corpus can be used for a study of chronological changes in modern Japanese. In addition, when compared with fiction data collected in the existing English corpora such as Brown and LOB, it can also be used for a comparative study of Japanese and English literature.

Keywords
日本語小説, 時系列データ収集, 機械翻訳, 形態素解析, 日英小説対照研究

1. はじめに

言語変化を観察しようとする場合, 小説は主要なデータ源の 1 つとなる。小説は, 不変の文学的テーマを追求しながらも, 時々の世相や風俗を敏感に反映するからである。英語について言えば, 小説は, 古くからコーパス内の独立したジャンルとして組み込まれてきた。1964 年に完成した Brown Corpus の場合, 小説は, 一般小説(サンプル数 29 本), 推理小説(24 本), 空想科学小説(6 本), 冒険小説(29 本), 恋愛小説(20 本), ユーモア小説(9 本)の 6 種のサブジャンルに区分され, 体系的な収集がなされている。その後, Brown の基準を踏襲する形で様々なパラレルコーパスが作られてきた(石川, 2021)。これにより, たとえば, Brown, Frown, Crown の小説データを比較することで 1961 年, 1991 年, 2009 年のアメリカ小説の言語変化が, LOB, FLOB, CLOB を比較することで同じ間隔におけるイギリス小説の言語変化が調査できる。

一方, 日本語の場合, 小説の言語変化を同じような形で検証することは容易ではない。2011 年公開の「現代日本語書き言葉均衡コーパス」(BCCWJ)には「書籍」ジャンルがあり, 日本十進分類法の「9 文学」に限定することで, 小説などを検索対象にすることはできるが, これには制約もあ

る。1 点目は，小説だけでなく，文学の研究書や図録などが含まれることである。2 点目は，海外小説の翻訳や古い作品の翻刻が含まれることである。3 点目は，年ごとのデータ量が統制されていないことである。BCCWJ の「短単位語数表」(Version 1.1)によると，「9 文学」のサンプル数は 1971 年が 2 本，1981 年が 16 本，1991 年が 206 本，2001 年が 595 本となり，年ごとの差は大きい。

筆者はかつて，英語小説との比較研究のため，Frown/ FLOB の標本抽出の基準年である 1991 年に限り，BCCWJ の「9 文学」に含まれる小説を手作業で抜き出し，Brown のサブジャンルを割り振って，小規模な日本語小説データセットの構築を試行した（石川，2015）。しかし，このデータには，(1)サブジャンルの判定が困難で分類の妥当性が保証されない，(2)BCCWJ を再構成したため独自の公開や活用ができない，(3)単一年のデータしかないため経年調査ができない，(4)日本語小説を集めただけでは英語小説との比較が困難である，といった問題があった。そこで，新たなコンセプトのもと，「1961-2021 日本語小説コーパス」（61-21JFIC）を開発することとした。本稿は 61-21JFIC の設計と構築の過程，また，収集データを用いた調査結果の一端を報告する。

2. 61-21JFIC の理念と構築過程

2.1 日本語小説テキストの収集

石川（2015）の 4 つの制約をふまえ，新しいコーパス開発では，(1')主観的なジャンル分類を廃し，Brown で言う「一般小説」のみを対象とする，(2')既存コーパスの再編集ではなく，新規にデータ収集を行う，(3')1961 年，1971 年，1981 年，1991 年，2001 年，2011 年，2021 年という 7 つの標本抽出ポイントを設定する，(4') 対応する年次の英米小説との語彙や表現の比較を可能にするため，英訳データを付与する，という 4 つの目標を掲げた。

まず，コーパスの枠母集団を決める準備として，Brown の「一般小説」に収録された作品の内容を調査し，それらの多くが日本で言う本格文学や純文学に類するものであることを確認した。次に，同様の日本語小説の収集の可能性を調べるため，神戸大学図書館において，書籍刊行年別に「日本文学」に分類されている作品の収録状況を調査した。その結果，たとえば 1961 年刊行のものは 250 件程度収蔵されているものの，大部分は研究書で，いわゆる小説作品は十分に収蔵されていないことがわかった。そこで，新たなアプローチとして，1960 年代から現在まで継続的に刊行されている文芸誌 3 種（『群像』，『新潮』，『文學界』）を枠母集団とし，上記の 7 か年のそれぞれ 1 月号に所収された小説（文芸評論・翻訳作品は除き，エッセイは含める）を収集対象に決めた。なお，1 月号が図書館に欠本の場合は古書で購入することとし，それでも入手不可の場合は図書館所蔵の同年の別月号からデータを取った。

データ抽出の開始点は，一律に作品冒頭部とした。ただ，文芸誌には，読み切り作品に加え，連載長編も多く掲載されているため，結果として，作品の長さやデータの抽出開始点についても一定の多様性が確保されたと言える。

サンプル数については，Brown の「一般小説」のサンプル数である 29 本を下回ることのないよう，各年につき，3 誌全体から 31 本を取ることとした。その際，作家の重複がなるべく起こらないよう調整して実際に取る作品を決定した。

サンプル長は一律 5,000 字とした。Brown のサンプル長が 2,000 語で，一般に英語 1 語が日本語 2~3 文字に相当するとされているためである。ただし，一部のデータについては，元の作品が短かったり，あるいは，作業時に読み取れなかった箇所を後から補填したりしたため，最終的な語数には若干のずれがある。

テキスト化については，当初，optical character reader（OCR）を用いた自動テキスト化を試みたが，古い雑誌は活字も不鮮明で，読み取り精度はきわめて低かった。そこで，すべてのサンプルについて，専門業者（作業者は 1 名に固定）に手作業で入力を依頼することとした。その際，(1)題名・作者名などは省略，(2)段落頭のアキは原稿通り，(3)文章段落頭以外の小見出し・番号・字下げ・地付き文字などはすべて詰めて入力，(4)空行（空改行）は 1 行分のみ入力，(5)括弧記号は全角で入力，(6)アルファベット略号は全角で入力，(7)横書きの欧文は半角で入力，(8)算用数字は 1 桁を全角，2 桁以上を半角で入力，(9)旧かな・旧漢字は新字体で打ち直して入力，(10)くの字点・踊り字（ゞゝゞヽ）はひらがな・かたかなで入力，(11)一部の環境依存文字・旧字体・正字・Unicode 文字は新字体・拡張新字体で入力（例：躯，掴など），(12)データは UTF8 処理，などの作業方針を指示した。なお，業者には文芸誌の現物のコピーデータを提供したが，コピーの問題で，見開き中央部が読み取れないものについては，それを除いて 5,000 字分を入力するよう依頼し，納品後，現物を確認しながら，筆者が不足箇所を手作業で補填入力した。

テキスト化されたサンプルは，すべて，国立国語研究所が提供する形態素解析システム「Web 茶まめ」上で処理した。解析辞書は国語研究所が開発した UniDic である。UniDic は意味を持つ最小単位で文字列を分割するため，揺れのない安定した結果が得られるが，たとえば，「図書館」が「図書」「館」に，「研究所」が「研究」「所」に，「していた」が「し（する）」「て」「い（いる）」「た」になるなど，一般に言う語より小さい単位で解析される。現時点では人手チェックは行っていない。

以上の過程を経て最終的に収集されたサンプルは，作家 156 名，作品 217 本，総語数約 70 万語となった。これらをコーパスに収録するにあたり，著作権者には必要な対応を取った。

2.2 英訳作成

構築コストの制約から，英訳については自動翻訳技術を使用することとし，ドイツの DeepL GmbH が提供する DeepL（DL）と，日本の情報通信研究機構（NICT）が提供する「みんなの自動翻訳@TexTra」（MN）という 2 種のニューラル機械翻訳（neural machine translation）で訳文を生成した。従来の自動翻訳は単語ベースの統計的機械翻訳（statistical machine translation）であったが，ニューラル機械翻訳では，あらかじめ膨大な対訳データを学習しておき，原文を数列で表現される特徴量（文の意味に相当）に変換した上で，その数値に合致する翻訳文を選び出して出力する。その際，複層的なニューラルネットによって原文の持つ特徴量を段階的に抽出・圧縮・加工していく（三竹，2018）。

DL は，ウェブ上の対訳データを集めた Linguee を基盤とする。特徴情報を抽出する畳み込み層（convolution layer）と，不要情報を除去するプーリング層（pooling layer）からなる畳み込みネットワーク（convolutional neural network）によって高い翻訳精度を実現しており，2020 年から

は日本語入出力にも対応している。一方，MN は NICT と総務省が 2017 年に立ち上げた「翻訳バンク」を基盤とする。同バンクには中央官庁・自治体・企業から大量の対訳データが提供されている。MN はユニバーサルコミュニケーション研究所多言語翻訳研究室で開発した自動翻訳エンジンをベースにしており，専門文献の翻訳にも強い。もっとも，ニューラル翻訳では原文の訳し漏れなどもしばしば発生する。今回は独立した 2 種の英訳データを集めることで，この問題に一定の対処を取った。英訳作成は 2021 年 4～8 月に実施し，最終的に生成されたデータは，DL が約 52 万語，MN が約 50 万語であった。現時点では人手による訳文の確認・修正は行っていない。

2.3 検索プラットフォームの開発

著作権処理を経て，現在，61-21JFIC のデータをオンラインで公開するための検索プラットフォームの開発に着手している。筆者の研究室ではすでに学習者コーパス ICNALE のための専用検索システム（ICNALE *Online*）を公開しており，これに改修を加え，日本語小説を日本語・英語（DL 訳文）・英語（MN 訳文）の 3 つのモードで検索できるようにする予定である。

新プラットフォームには，KWIC 検索，コロケーション検索，単語頻度検索（語彙表出力），特徴語検索などの基本機能のほか，語や表現の出現頻度の時系列的な変化を可視化するグラフ出力機能を実装する。これらの機能によって，各年代の特徴語を自動抽出したり，指定した語や表現の頻度変化を調べたりすることが可能になる。

3. 61-21JFIC でわかること

3.1 日本語の時系列分析

61-21JFIC により，ジャンルを厳格に統制した上で現代日本語の時系列変化の調査が可能になる。図 1-4 は，語種（和語・漢語・外来語・混種語），句読法（句点・読点），高頻度品詞（助詞ほか 4 種），低頻度品詞（感動詞ほか 7 種）の 10 万語あたり頻度の変化を示したものである。1961 年と 2021 年を比較して 10%以上増加したのは外来語（+140%）と漢語（+12%）の 2 つで，減少したのは接続詞（-31%），代名詞（-22%），連体詞（-15%），形容詞（-12%）である。このうち 60 年間にわたって線形的に頻度が変化しているのは外来語のみで，全体で見れば語種や品詞の頻度は安定していると言える。

一方，具体的な語の中身に踏み込むと当然ながら年代ごとに差異が認められる。こうした差異は，安定的に出現すると思われる助詞にも表れている。図 5 は年代を第 1 アイテム，高頻度助詞（の・に・を・が・と・で・から）頻度を第 2 アイテムとして対応分析を実施した結果である。1961 年～1991 年が第 1 軸（寄与率 75.9%）上で左側に，2001 年～2021 年が右側にそれぞれ固まっており，1991 年と 2001 年の間に日本語小説中の助詞選択にある種の変化が生じている可能性が示される。このデータに限って言えば，1960～90 年代の小説では，連体節（「の」）や起点表現（「から」）が多かったのに対し，2000 年代以降になると，並列表現（「と」）や，主格や目的格といった格標識（「が」「を」「に」）が顕著になっているものと思われる。

図 1~4 語種・句読法・高頻度品詞・低頻度品詞の頻度変化

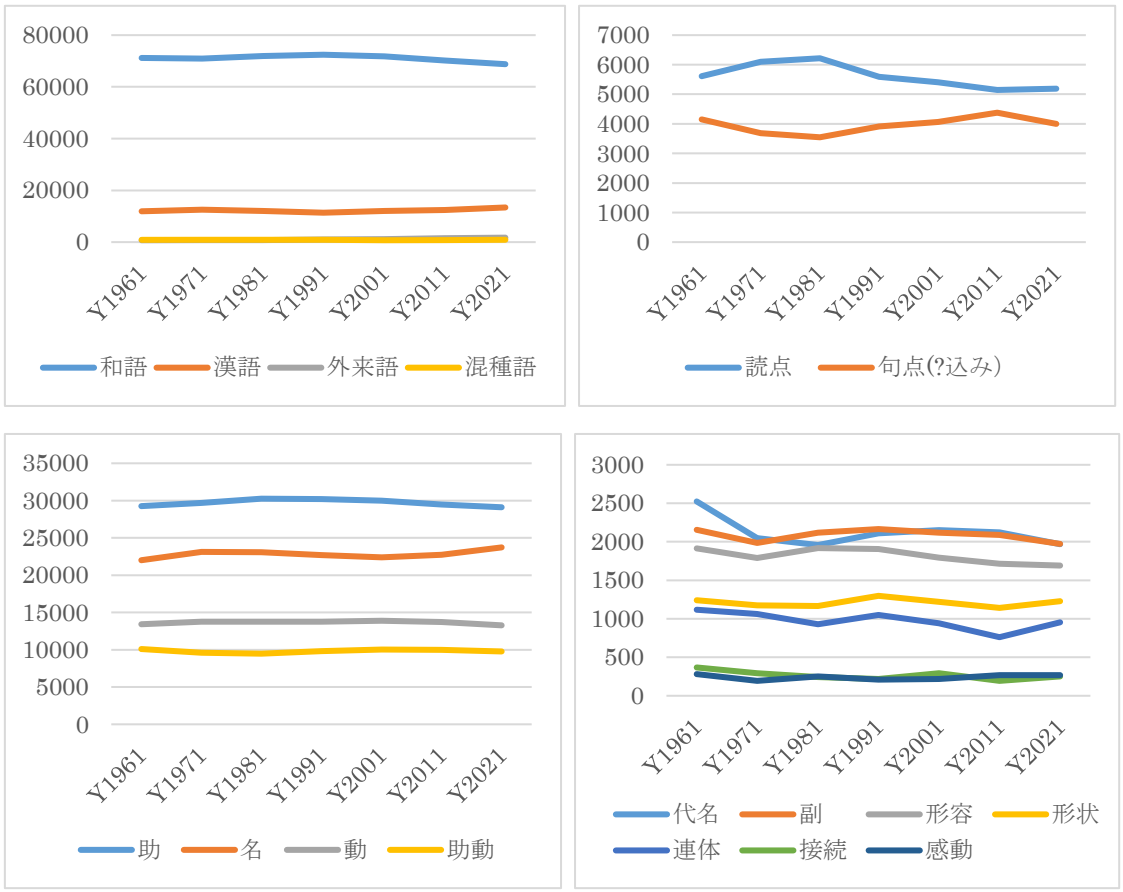
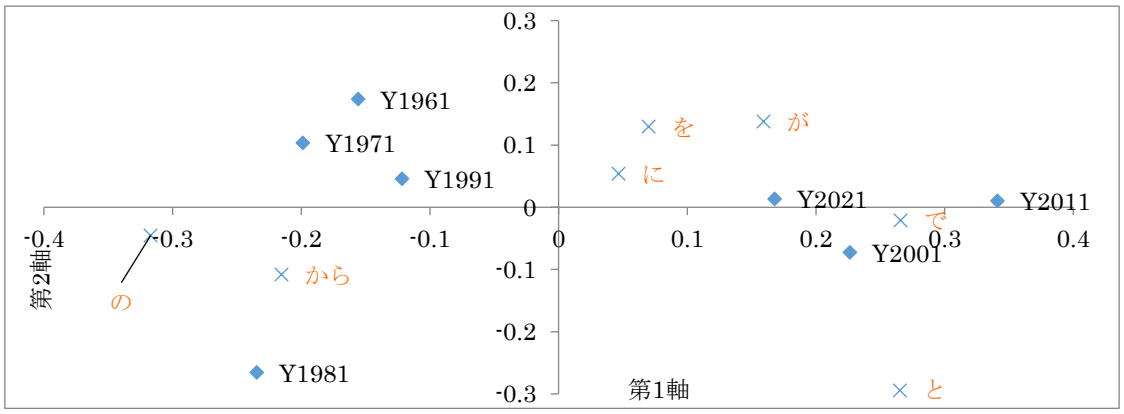


図 5 年代×助詞の頻度表に対する対応分析の結果



3.2 英語小説との対照分析

61-21JFIC はまた，英訳データと既存の英語コーパスの小説データを比較することで，時代要因を統制した日英小説の対照分析にも使える。表 1 は 1961 年に対象年次を限定し，日本の小説（DL 英訳），Brown 収録のアメリカ小説，LOB 収録のイギリス小説の各々をそれらの総体データ

と比較し，特徴語を抽出した結果である。統計量が 15 以上であれば 0.01%水準で有意である。

表 1 1961 年の日本語・アメリカ・イギリス小説の特徴語（対数尤度比統計量：LL）

順位	日本		アメリカ		イギリス	
1	i	146.64	he	57.6	her	77.13
2	my	127.42	his	36.66	she	61.31
3	s	45.35	h	30.28	julian	30.22
4	m	43	church	28.02	said	24.48
5	is	36.88	scotty	27.14	you	23.64
6	a	29.2	john	26.15	love	20.36
7	if	25.49	him	24.88	vera	20.36
8	tokyo	24.43	alex	23.83	queen	18.61
9	t	23.94	winston	23.83	celia	18.39
10	wife	21.6	watson	22.51	joe	18.08

ここで詳細な内容分析に踏み込む余裕はないが，今回のデータに限って言えば，1961 年時点において，日本の小説の内容が 1 人称中心で（I, my），会話が多く（s, m, t などの縮約辞），仮定文（if）が散見されるのに対し，アメリカ小説には男性 3 人称（he, his, him）が多く，イギリス小説には 2 人称（you）や女性 3 人称（her, she），また，発話動詞（said）が多いことがうかがえる。

4. まとめ

以上で概観したように，61-21JFIC は，日本語の時系列分析の資料として，また，時代要因を統制した日米英の小説比較研究の資料として，様々な活用が考えられる。後者について言えば，言語や内容に大きな影響を及ぼす時代要因を完全に統制できる利点は大きく，同時代の各国小説に見られる「マインドスケープ」などの計量的比較も可能になるだろう。こうしたアプローチは，主観的な批評にとどまっていた従来の比較文学研究に新しい展開をもたらすものと言える。

謝辞

本研究は JSPS 科研費（挑戦的研究（萌芽））20K20699「言語から見た日米マインドスケープ比較：データサイエンス志向型小説研究の試行」の助成を受けている。

引用文献

石川慎一郎(2015)「FROWN/FLOB Corpus および BCCWJ データの再構成に基づく英日対照言語研究用小説テキストデータセットの構築の試み：English-Japanese Modern Fiction Corpus (EJ-MoFic)の概要」『統計数理研究所共同研究リポート』*340*, 1-18.

石川慎一郎（2021）『ベーシックコーパス言語学』第 2 版．ひつじ書房．

三竹保宏（2018）「Deep Learning による AI 機械翻訳のイノベーション」『ビジネスコミュニケーション』*55*(8), 11.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

中英語期の補部と 2 人称代名詞の構文関係
―ICAMET による分析―

泉類　尚貴（慶應義塾大学 大学院生）
n.senrui@gmail.com

Constructions with Complementation and the Second-Person Pronouns in Middle English
―Analysis with ICAMET―

SENRUI Naoki (Keio University, Graduate Student)

Abstract

This study attempts to examine the constructions with complementation (*that*-clause or infinitives) and the second-person pronouns in Middle English with the analysis of ICAMET. The construction treated in this paper is directive performatives, which have existed throughout the history of the English language. I consider the relationship between the directive speech act verbs, the second-person pronouns and the complementation. We cannot observe significant features among them in terms of quantity except some verbs. Besides, with the COCOA tags attached in ICAMET, it is discussed whether the sociolinguistic factors such as authors and genres influence the choice of complementation. In some works, *that*-cl occurs with vocatives such as *sir*, which implies that the finite complement represents a polite function when one can choose from among different types of complementation.

Keywords
Explicit Performatives, Directives, Late Middle English,
the Second-Person Pronouns, Sociolinguistics

1. はじめに

英語史における変化の一つに，補部の変化がある。定型節から非定型節への変化は，中英語期にその萌芽が見られる（Los, 2005; Manabe, 1989）。補部と意味の関係について，Rohdenburg（1995）によれば，非定型節のほうが coercive force が強いことが示されている。Coercive force の強さは，*command* をはじめとする指令動詞の分析から提示された。一方で，同一の動詞が異なる補部を従える例も見られる。本研究では，補部の変化が起こりだした時代である中英語期に焦点をあてて，指令動詞の現れる構文の一つである，明示的遂行文における 2 人称代名詞の使用区分（*thou* 系か *ye* 系か）と補部の形式（*that* 節か不定詞か）について，Innsbruck

Computer Archive of Machine-Readable English Texts（ICAMET）を中心に収集したデータを示し，分析の可能性を示す。

2. 先行研究

英語史において，動詞補部は *that* 節から不定詞，不定詞から動名詞へと変化する傾向にあることが Rohdenburg（1995 ほか）により示された。Rohdenburg（1995）は，Defoe や Swift をはじめとする 16 世紀から 18 世紀に書かれた作品を資料として，補部と coercive force の関係について，*command* や *order* をはじめとする，manipulative verbs を対象に調査を行った。調査を通して，定型節（*that* 節）と非定型節（不定詞）を比較した場合，定型節のほうが coercive force が弱いことを示した。*that* 節から不定詞への補部変化は，古英語期と初期中英語期を比較した Los（2005）や Manabe（1989）の調査により，初期中英語期からその変化が起こっていたことが示されている。

Coercive force は，ポライトネスと関与する可能性がある。ポライトネスは，Brown and Levinson（1987）によるフェイスをもととした理論をはじめとして，数多くの理論が構築されている。現代的なポライトネスの概念は，古英語期にはほとんど見られず（Kohnen, 2008），中英語期に勃興したことが，Jucker（2020）をはじめとして論じられている。

行為指示動詞と結びつく 2 人称代名詞について，*pray*, *beseech* のコンコーダンス冒頭 100 例を基に，Shakespeare の悲劇を対象に調査した Brown and Gilman（1989）によって，*beseech* はその大半が敬称の Y 系と共起していることが示されている。

補部は，ジャンルをはじめとする社会言語学的要因とも関わる。Mair（2002）は，補部が不定詞であるか，動名詞であるかの英米差，ならびにジャンルの差について，現代英語の *start* と *begin* を対象として調査を行った。英米差の観点では，アメリカ英語のほうが動名詞を好み，ジャンルの観点では，とくに *start* に後続する補部について，会話では動名詞が好まれ，記述文では不定詞が好まれることを示した。

3. リサーチデザイン

3.1 研究目的と研究設問

先行研究は主に近代英語以降を対象とした調査であるが，本研究では，現代的なポライトネスが勃興したとされる中英語期に焦点をあてる。とりわけ，ポライトネスがもっとも見られるスピーチ・アクトの一つである Directives（行為指示）に注目する。形式は，遂行文，かつ，話し手と聞き手の関係が明らかである，1 人称主語＋行為指示動詞＋2 人称目的語，あるいは 1 人称主語＋行為指示動詞＋補部標識 *that*＋2 人称主語に限定して調査を行う。遂行文に注目するのは，Kohnen（2008）らによって，古英語・中英語期から用いられていた形式と示されているためである。コーパスは，ICAMET を用いる。これは，600 万語を超える資料を収録しており，中英語のコーパスとしては最大規模の資料の一つであること，ジャンルをはじめとする COCOA タグが付与されているためである。

補部の選択については，中英語期に関して，ポライトネスを示す手段の一つとして用いられた，2 人称代名詞の敬称（Y 系）と親称（T 系）の選択と関与する可能性がある。中英語期は，2 人称代名詞が技巧的に用いられている例も見られ，数多くの研究がなされてきたが，遂行文における構文関係について言及した研究は見られない。また，Traugott（2012）の指摘や，Helsinki Corpus，ICAMET に付与された COCOA タグの示すように，中英語期は新たなジャンルが勃興した時代でもある。

これらの構文の調査を通して，主に ICAMET を資料に，次の問題について，議論を行う。1 点目は，2 人称代名詞と補部の関係について考察する。続いて，社会言語学的要因に注目する。2 点目として，ジャンルと補部の関係性について考察する。3 点目は，作者・作品と補部の関係性について考察する。先行研究に照らせば，不定詞のほうが coercive force が高いことから，T 系と結びつく傾向にあると推察できる。ジャンルについて，古い形式を好むであろうロマンスをはじめとするジャンルでは，古くから存在した形式である *that* 節が，一方で書簡をはじめとする，口語を反映しやすいジャンルでは，新しい形式である不定詞が選択される可能性があると予想できる。また，個々の作家や作品が補部の選択に与えた影響についても，データを収集する。

3.2 データ

本論では，Manfred Markus 氏により構築された ICAMET を用いる。加えて，これまでに PPCME2 や Helsinki Corpus を主たる対象とし，ポライトネスにもっとも配慮する必要があると考えられる 1 人称主語，ならびに 2 人称が目的語，あるいは *that* 節の主語が 2 人称の場合に生起する構文の環境において，行為指示を表す動詞について調査し，*pray*，*beseech* をはじめとする 18 の動詞を収集した。これらの動詞を，DSAVs（Directive Speech Act Verbs）と名付ける。

なお，中英語期に見られる，本動詞としての *will* を除外するのは，directives にも，他のスピーチ・アクトである commissives にもなりうるため，曖昧性が生じるからである。

個々の作品について，ICAMET に含まれている *The Paston Letters* は，書き手と受け手の情報をはじめとするタグが付与された PCEEC（The Parsed Corpus of Early English Correspondence）を基にして，分析を行う方針であるため，本研究では除外する。

3.3 手法

3.3.1 分析の手順

KWIC Concordance を利用し，ICAMET から DSAVs を検索する。コーパスに付けられている COCOA タグを利用し，研究設問に答えるため，<A>（著者），<B>（タイトル），<C>（世紀），<T>（ジャンル）を出力する。

出力されたデータに，代名詞については 2 人称代名詞が T 系か Y 系か，補部については *that* 節か不定詞かについてタグ付けを行う。不定詞には，中英語期に見られる不定詞のマーカーである，*for to* をはじめとするマーカーを含む。なお，本研究において，補部が名詞句や前置詞句である用例はデータから除外する。

表 1 構文別の DSAVs と補部

	advise		ask		beseech		bid		charge		command	
	Y	T	Y	T	Y	T	Y	T	Y	T	Y	T
that 節	2 (50%)	0	3 (75%)	1 (100%)	48 (62%)	26 (63%)	3 (75%)	5 (71%)	20 (90%)	13 (93%)	23 (74%)	9 (82%)
不定詞	2 (50%)	0	1 (25%)	0	29 (38%)	15 (37%)	1 (25%)	2 (29%)	2 (10%)	1 (7%)	8 (26%)	2 (18%)

	counsel		desire		exhort		pray		rede		require	
	Y	T	Y	T	Y	T	Y	T	Y	T	Y	T
that 節	26 (79%)	11 (58%)	1 (50%)	0	1 (50%)	1 (100%)	240 (75%)	39 (76%)	1 (50%)	15 (100%)	28 (93%)	3 (50%)
不定詞	7 (21%)	8 (42%)	1 (50%)	0	1 (50%)	0	90 (25%)	12 (24%)	1 (50%)	0	2 (7%)	3 (50%)

4. 結果と考察

4.1 RQ1: 2 人称代名詞と補部の構文関係について

DSAVs のうち，*that* 節，不定詞の両方を補部とする動詞は，表 1 に数値とともに示した 12 の動詞に限られた。なお，綴り字は現代英語の綴りで表記している。このうち，*advise* や *ask* をはじめとする動詞は，頻度が低いため，具体的な分析が難しい。また，*rede* については，不定詞を補部として取る例は 1 例に過ぎない。

頻度の高い DSAVs である，*beseech*, *charge*, *command*, *counsel*, *pray*, *require* について，いくつかの結果を示す。生頻度をもとに割合を計算したところ，たとえば，*pray* について，Y 系，T 系ともに，*that* 節ならびに不定詞を従える割合は，前者がおおよそ 75%，後者がおおよそ 25%で，2 人称代名詞による頻度の差異は観察されない。また，数値をカイ二乗検定にかけたところ，$p. < .05$ の有意差は，*beseech*, *pray*, *charge*, *command*, *counsel* の各動詞について，見られなかった。すなわち，2 人称代名詞と補部に強い相関関係は見られない。2 人称代名詞の選択は，ポライトネスを反映した言語形式であるが，少なくとも数値上は，補部にポライトネスを示す機能は見られないようである。

一方，*require* については，$p. < .05$ の有意差が見られることから，研究設問において予想したとおりの結果である，T 系の代名詞のほうが不定詞を補部として従えることが分かった。不定詞のほうが，coercive force が高いことは，Rohdenburg（1995）による近代英語期の分析から示されている。中英語期において，Y 系と比較して，より強く指示できる場面が多いと考えられる T 系と，不定詞の共起率が高いことから，不定詞により強い coercive force があることは，一部の動詞には適応されると見なせる。

4.2 RQ2: ジャンルと補部の関係性について

ジャンル・作者と補部の関係について，いくつかの調査結果を示す。ジャンルについては，ICAMET の COCOA タグに原則として従う。

ICAMET には，32 のジャンルに分類したタグが付いている。Sermons では，すべての補部が不定詞である。Sermons は，聞き手のフェイスを侵害しないように，配慮する必要がないジャンルであると見なせる（Kohnen 2008: 304）。15 世紀の sermons を調査した Kohnen（2008）によれば，フェイス侵害を避けるために用いられうる間接言語行為は 1 例に留まることが示されている。すなわち，フェイス侵害を考慮する必要がないため，変化の過渡期にあった新しい形式を採用したと見なせる。

書簡は，ICAMET に含まれるジャンルの中でも，もっとも日常の言語が反映されたジャンルであると見なせる。例として，動詞 *beseech* を検討する。118 例現れる用例のうち，書簡で現れるのは 22 例である。代名詞については，後期中英語期の書簡においては Y 系が一般的であり，本構文で共起した代名詞も，すべて Y 系である。補部は，*that* 節が見られず，すべて不定詞である。

4.3 RQ3: 作者・作品と補部の関係性について

続いて，例えば，Chaucer の直筆の写本は現存していないが，広い意味での作者について言及する。全ての DSAVs を調査したところ，いくつかの作品を翻訳した Caxton の作品においては，*that* 節，ならびに不定詞の両方を補部として選択している例が複数見られる。一方，Chaucer の *The Canterbury Tales* に含まれる話の一つである，'The Tale of Melibeus' は *that*-cl のみしか従えていない。*counsel* の用法について，Richard Rolle では不定詞しか補部として共起しない（4 例）。一方，*Prose Life of Alexander* では，すべてが *that* 節と共起する（4 例）。*require* の用法について，T 系と Y 系と共起する例を合計して現れる 5 例は，*Merlin*（2 例），*Le Morte Darthur*（2 例），ならびに *Melusine*（1 例）の 3 作品にのみ現れる。ただし，これらの 3 作品には，*that* 節を補部として取る例も見られる。用例数は *Merlin*（9 例），*Le Morte Darthur*（4 例），*Melusine*（5 例）。これらの 3 作品については，*that* 節と不定詞を，書き手が「選択」していたと想定される。例として，*Le Morte Darthur* において，*that* 節をとる例は，いずれも *sir* を含む呼びかけ語と共起している。すなわち，*sir* を用いた場合，書き手は *that* 節を用いたことになる。この一方で，不定詞をとる例は，*sir* との共起は見られない。

5. まとめ

本研究では，中英語期の最大規模のコーパスである ICAMET を用いて，中英語期の行為指示を表わす遂行文を対象として，二人称代名詞と補部（*that* 節か，不定詞か）の構文関係についての調査を行った。加えて，COCOA タグを利用し，ジャンルや作者をはじめとする社会言語学的調査を行った。二人称代名詞と補部の構文関係について，多くの動詞においては，T 系，Y 系と補部の強い相関性は見られなかった。一方，いくつかの動詞については，不定詞と T 系に強い構文関係が観察された。この事実は，Rohdenburg の示した coercive force が，中英語期についても適

応できる可能性を示している。ジャンルについて，sermons では不定詞補部が大半であるといった傾向が見られた。また，補部を作者が選択できる場合，*that* 節と *sir* をはじめとする呼びかけ語が共起しており，ポライトネスが高い形式である可能性を示唆した。

本調査では， 2 人称代名詞と補部の構文関係について，強い差が見られなかった動詞が多かった。しかし，差が見られなかったという事実は，選択することの可能性の高さ，あるいは固定していた可能性を示しているとも見做せる。そこで，今後の課題として，場面に注目した，社会語用論的調査と考察を行うことをあげる。また，数値の統計的処理についても，検討していく指針である。

謝辞

本研究は，潮田記念基金による慶應義塾博士課程学生研究支援プログラムの助成を受けたものです。

引用文献

ICAMET= Markus, Manfred, ed. *The Middle English Prose Corpus of the ICAMET*. Innsbruck: U of Innsbruck, 2003.

Brown, Penelope and Stephen C. Levinson. *Politeness: Some Universals in Language Usage.* Cambridge: CUP, 1987.

Brown, Roger and Albert Gilman. 'Politeness Theory and Shakespeare's Four Major Tragedies.' *Language in Society*, 18:2 (1989), 159-212.

Jucker, Andreas H. *Politeness in the History of English: From the Middle Ages to the Present Day.* Cambridge: CUP, 2020.

Kohnen, Thomas. 'Tracing Directives through Texts and Time: Towards a Methodology of a Corpus-based Diachronic Speech-Act Analysis.' *Speech Acts in the History of English*. Ed. Andreas H. Jucker and Irma Taavitsainen. Amsterdam: John Benjamins, 2008. 295-310.

Los, Bettelou. *The Rise of the To-Infinitive.* Oxford: OUP, 2005.

Mair, Christian. 'Three Changing Patterns of Verb Complementation in Late Modern English. A Real-Time Study based on Matching Text Corpora.' *English Language and Linguistics*, 6 (2002), 105–131.

Manabe, Kazumi. *The Syntactic and Stylistic Development of the Infinitive in Middle English.* Kyushu University Press, 1989.

Rohdenburg, G. 'On the Replacement of Finite Complement Clauses by Infinitives in English.' *English Studies*, 76:4 (1995), 367-388.

Traugott, Elizabeth Closs. 'Pragmatics and Discourse.' *English Historical Linguistics: An International Handbook. Vol. 1.* Ed. Alexander Bergs and Laurel J. Brinton. Berlin: Mouton de Gruyter, 2012. 466-480.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

Verification of the Effectiveness of 20 Months of Speaking Lessons for High School Learners —An Analysis of Fluency on the Aptis Speaking Test—

TIKHONENKO Maksim
(Tokyo University of Foreign Studies, Graduate Student)
maximtikhonenko72@gmail.com

MOCHIZUKI Keiko
(Tokyo University of Foreign Studies)
mkeiko@tufs.ac.jp

Abstract

This presentation presents a 20-month longitudinal study on the development of speaking ability observed in Japanese high school learners of English who participated in online speaking lessons. Students were divided into two groups: an experimental group who had taken 20 lessons, and a control group who had only taken 3 lessons. Speaking data recorded during the time when students took the Aptis speaking test was transcribed, and the duration of pauses and speaking time was measured. Then the transcribed data was analyzed from the point of view of fluency and complexity. The analysis showed that fluency developed significantly, especially in terms of speech rate and pause time to speaking time ratio. On the other hand, complexity did not develop as much.

Keywords

Longitudinal Learner Corpus, Fluency, Complexity, High School Learners, Aptis Speaking Test

1. Introduction

To understand how lessons of spoken English can help Japanese high school students in acquiring speaking skills, a research team from Tokyo University of Foreign Studies organized an educational project in collaboration with Sankei Human Learning, Lingua House, and two high schools.

High school students took 20 lessons in spoken English and had free conversation with teachers from the Philippines using Zoom. Students also wrote essays on the topic of each lesson, and filled in surveys to evaluate their progress after each lesson and record what was most difficult for them. Each lesson lasted about 25 minutes, and 32

students started the lessons in the first year, November 2018, and completed the 20th lesson in the 3rd year, July 2020.

To understand the contribution of the lessons, it was proposed that a 3-lesson control group consisting of 22 students who only took 3 lessons (lessons 15-17) in the 2nd year, January-March 2020, also participate in the project. Both groups used a textbook written by members of the research team. Two research questions were posed:

Research question 1: Do online lessons of spoken English held monthly for 20 months positively affect the development of fluency and complexity in Japanese high-school learners of English, and which aspects?

Research question 2: How different are the fluency and complexity of English spoken by learners who have been taking the online lessons for 16 months in comparison to the English of those who have not taken such lessons?

Data was obtained from video recordings of 25-minute lessons. Video and audio data were transcribed manually, and the transcribed texts of selected learners were divided into units of speech. The ELAN software was used to measure the free conversation time in each lesson.

2. The APTIS test

Data was obtained from video recordings of 25-minute lessons. At the end of the project, all students took the APTIS speaking test in September 2020 in the 3rd year.

The APTIS test is a test developed by the British Council, based on the Common European Framework of Reference. It consists of 4 parts:

Part 1 (A1/A2 level), in which students answer three simple questions about their background and experiences. 30 seconds to response to each question.

Part 2 (B1 level), in which students describe a photograph and answer two questions related to it. 45 seconds to response to each question.

Part 3 (B1 level), in which students describe and compare two photographs, and answer two questions. 45 seconds to response to each question.

Part 4 (B2 level), in which students answer three more abstract questions.

1 minute to prepare a response and 2 minutes to answer all questions in one go.

Students' speaking and grammar & vocabulary skills were also evaluated by the British Council and average scores were calculated for the two groups (Table 1).

The difference in speaking score was greater than that of the grammar & vocabulary score, which may indicate that the main effect of the lessons was on speaking skills and not on language knowledge.

Table 1 *Average APTIS scores of the 20-lesson and 3-lesson groups*

	Speaking	Grammar & Vocabulary
20-month group	32.19	36.38
3-lesson group	28.27	34.41
Difference in score	3.92	1.97

3. Data elicitation and method of analysis

8 learners were selected in each group (the 20 lessons group and the 3 lessons group) to analyze fluency and complexity parameters in the APTIS speaking data and compare them between the two groups.

Speaking data was recorded and manually transcribed using ELAN. Speaking time and pause time were measured. Next, the texts were divided into AS-units, and analyzed for fluency and complexity using Microsoft Excel.

An AS-Unit is a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either (Foster et al. 2000).

The following parameters were measured to evaluate fluency:

a) Speech rate: words/min;

b) Ratio of pause time to speaking time: pause time/speaking time;

c) Number of pauses per minute: number of pauses/speaking time;

d) Ratio of fillers to unpruned words: number of fillers / number of words including disfluencies;

e) Ratio of repetitions to unpruned words: number of repetitions / number of words including disfluencies;

f) Ratio of self-corrections to unpruned words: number of self-corrections / number of words including disfluencies.

For complexity, the following parameters were used:

g) Mean utterance length: number of words / number of AS-Unit;

h) Overall number of words;

i) Ratio of subordinate clauses to AS-Units: number of subordinate clauses / number of AS-Units.

4. Results

4.1 Fluency

Fluency is a complex property of L2 performance related to speed of speech, pause phenomena, and the impression a speaker leaves on a listener (Segalowitz, 2010).

Figure 1 *Pruned Speech Rate Aptis Speaking Test*

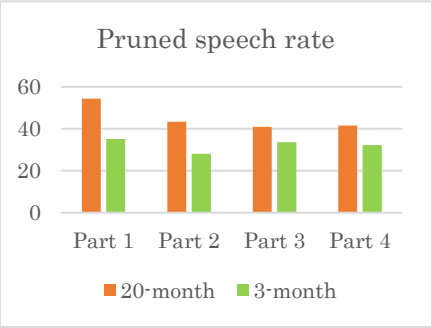


Figure 1 shows that the 20 lessons group displays a higher pruned speech rate than the 3 lessons group. In part 1, which corresponds to CEFR A1/A2 levels, the 20 lessons group showed the best performance, but in other parts their performance was lower and comparatively did not change much. In contrast, the performance of the 3 lessons group clearly improved between part 2 and part 3, which can be explained by the fact that to describe two pictures in part 3, the learners used simpler constructions with a higher speech rate.

Figure 2 *Ratio of pause time to speaking time*

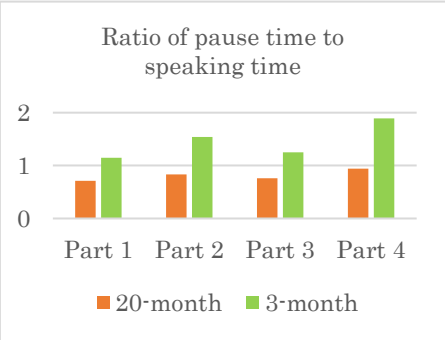


Figure 2 shows that the 20-lesson group produced fewer long pauses than the 3 lessons group. Their ratio of pause time to speaking time was below 1 in all tasks, which indicates that they spoke more than they kept silent. In contrast, the 3-lesson group produced more silent pauses than speech in all tasks.

Figure 3 *Number of pauses per minute*

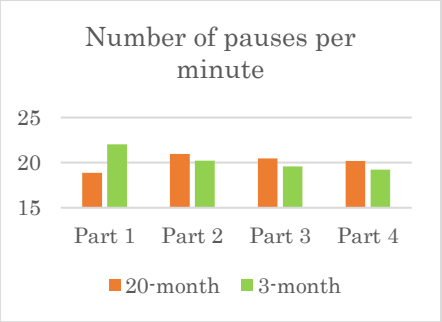


Figure 3 shows that the number of pauses in both groups was almost the same except for task 1, where the 20-lesson group slightly outperformed the 3-lesson group. This means both groups tended to produce the same number of pauses in more difficult tasks, but the length of pauses was significantly shorter for the 20-lesson group.

Figure 4 *Ratio of fillers to number of unpruned words*

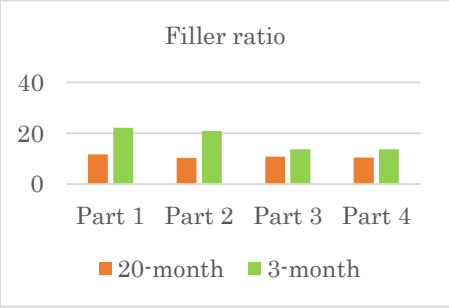
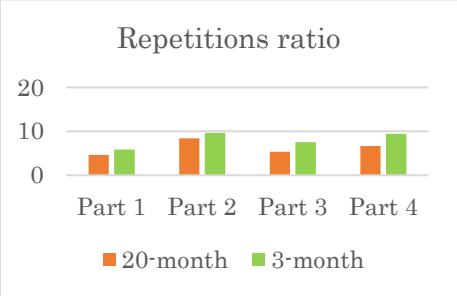


Figure 4 shows that the filler ratio is higher among the 3-month group in the first half of the test. That illustrates that the 3-month group produced more filled pauses, which contradicts the pattern observed during dialogues with teachers when both groups showed the same performance. This may indicate that the effects of taking lessons for 20 months showed up in a more predictable environment, where students were more

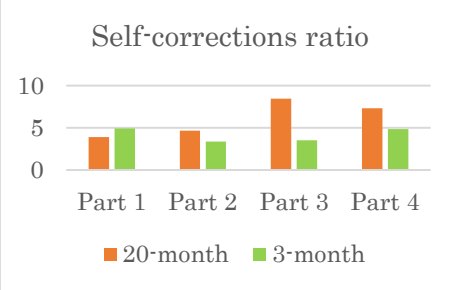
relaxed. Also notable is the fact that the performance of the 3-month group improved during the test, and they produced less filled pauses in the second half of the test.

Figure 5 *Ratio of repetitions to number of unpruned words*



To a lesser extent one can observe the same result in repetitions ratio, as figure 5 shows.

Figure 6 *Ratio of self-corrections to number of unpruned words*



However, the self-corrections ratio did not show clear correlations, as figure 6 shows.

4.2 Complexity

Figure 7 *Mean utterance length*

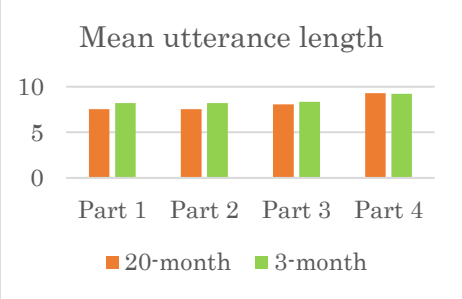
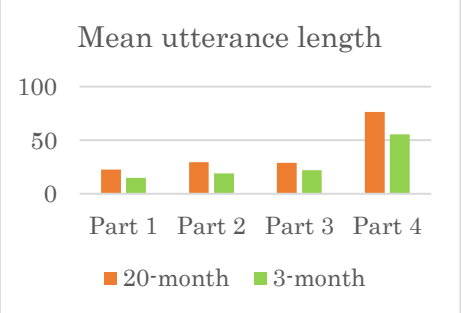


Figure 7 shows mean utterance length among the two groups. It is relatively similar for both groups, which may indicate that the lessons did not have any influence on syntactic complexity.

Figure 8 *Overall number of words*



However, the 20-month group tended to produce more words, as illustrated in figure 8. This indicates that while word production rate and therefore fluency increased, students did not gain the ability to build more elaborated utterances.

Finally, analysis of the mean subordinate clauses ratio was conducted. Subordinate clauses were quite rare except in part 4, so it was decided to analyze subordinate clauses only in

this part. Analysis shows that the mean ratio of subordinate clauses to AS-units was 0.48 for the 20-month group and 0.57 for the 3-month group. However, this does not illustrate that the speech of the 3-month group was more elaborated. Absolute numbers of subordinate clauses were between 3 to 5 in both groups, and the types of subordinate clauses were rather limited: they were introduced by the words "I think" and such conjunctions as "when" or "where". That means students in both groups used a similar limited inventory of syntactic means to express themselves, and the 20-month group did not tend to use more subordinate clauses than were minimally needed. But as the number of words and AS-Units in their responses were higher, the subordinate clause ratio ended up being lower than that of the 3-month group.

5. Conclusion

The comparative analysis of fluency and complexity of the two groups of students showed that 20 months of English conversational lessons had positive effects on students' fluency. affected, and effects were observed the most in speed and breakdown fluency.

All three components of fluency were affected, and effects were observed the most in speed and breakdown fluency. However, the effects were less evident in repair fluency, with less significant differences in the ratios of fillers and repetitions, and an absence of any correlations in the ratio of self-correction. At the same time, lessons did not have any significant effect on the syntactic complexity of students' English.

Therefore, we may conclude that students who had taken additional lessons of spoken English once in a month became more confident and could produce more words with less hesitation, but their English did not become more elaborate, at least syntactically.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP17H02357 and JP20H01278.

Bibliography

Foster, P., A. Tonkyn, G.Wigglesworth, (2000). "Measuring Spoken Language: A Unit for All Reasons", Applied Linguistics . Volume 21, Issue 3. 354-375. Oxford: Oxford University Press.

Segalowitz, N. (2010). Cognitive Bases of Second Language Fluency. New York: Routledge.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 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)

仁科 恭徳(神戸学院大学)
ynishina@gc.kobegakuin.ac.jp

赤瀬川 史朗(Lago 言語研究所)
lagoinst@gmail.com

Toward the Development of "Parallel Link (tentative name)" (Ver. 1.0), an Online Search Tool for Japanese-English and English-Japanese Parallel Corpora (Interim Report)

NISHINA Yasunori (Kobe Gakuin University)
AKASEGAWA Shiro (Lago Institute of Language)

Abstract

In this presentation, we will give an interim report on the progress of Parallel Link (Ver. 1.0), an online analysis tool for Japanese-English and English-Japanese parallel corpora that we are currently developing. First, we will review the Japanese-English and English-Japanese parallel corpora and analysis tools that have been developed to date. Next, we will report on the progress of Parallel Link (Ver. 1.0) under development, especially focusing on rebuilding and refurbishing the existing parallel corpora.

Keywords

パラレルコーパス, ツール開発, オンライン, 日英・英日翻訳

1. はじめに

本発表では, はじめに, 現在までに公開された日英・英日パラレルコーパス, 開発されたコンコーダンサーなどの検索ツール, 日英・英日パラレルコーパスを活用した研究を網羅的に振り返り, 今後期待される方向性に関してまとめる。続いて, これからのパラレルコーパス研究の方向性を示すべく目下開発中の日英・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)の開発経緯と進捗状況を報告する。可能であれば, 同ツールに実装予定の機能や, 同ツールを用いた想定されるケーススタディなど, 開発後の展望も具体的に示したい。

なお, 本プロシーディングスでは, 紙幅に限りがあることから, 同ツールに搭載予定のパラレルコーパスに関して簡単に触れる。そして, それらのフォーマットを統一するために施したテキスト処理を含む一連の作業内容を示す。

2. 『(仮称) パラレルリンク』(Ver.1.0) の開発に向けて

これまでの日英・英日パラレルコーパス（研究）の状況を受け，現在までに公開されている日英・英日パラレルコーパスを網羅的に串刺し検索できるオンライン検索ツール『(仮称) パラレルリンク』(Ver.1.0) を目下開発中である。本節ではそのプロトタイプとなる Ver.1.0 に関して，主に搭載予定のコーパスとテキスト処理など一連の整備作業について報告する。

2.1. 『(仮称) パラレルリンク』(Ver.1.0) に搭載予定のパラレルコーパスについて

本検索ツール (Ver.1.0) の開発に先立ち，既存のパラレルコーパスのデータを整備した。現時点で，対象とした日英・英日パラレルコーパスは「日英サブタイトルコーパス Japanese-English Subtitle Corpus (JESC)」(https://nlp.stanford.edu/projects/jesc/index_ja.html),「日英法令対訳コーパス (LAW)」(http://www.phontron.com/jaen-law/index-ja.html),「大規模オープンソース日英対訳コーパス (OPENSOURCE)」(https://www2.nict.go.jp/astrec-att/member/mutiyama/manual/index-ja.html),「ロイター日英記事の対応付けデータ (REUTERS)」(https://www2.nict.go.jp/astrec-att/member/mutiyama/jea/reuters/index.html),「SCoRE 用例コーパス (SCoRE)」(http://www.score-corpus.org/),「日英対訳文対応付けデータ (TAIYAKU)」(http://www2.nict.go.jp/astrec-att/member/mutiyama/align/index.html),「Tatoeba 日英対訳コーパス (TATOEBA)」(https://tatoeba.org/),「TED Talk 日英コーパス (TED) (ただし，センテンス単位ではない *字幕ファイル(ssa 形式ファイル)から作成)」(https://amara.org/en/teams/ted/videos),「Wikipedia 日英京都関連文書対訳コーパス (WIKIPEDIA)」(https://alaginrc.nict.go.jp/WikiCorpus/)の計 9 種である[1]。

表 1 既存パラレルコーパスの対訳対数と語数について

コーパス名	対訳対	語数（日本語）	語数（英語）
JESC	330,102	2,736,837	2,222,329
LAW	262,448	9,264,891	9,508,555
OPENSOURCE	505,780	6,927,281	5,018,603
REUTERS	70,120	2,068,681	1,740,428
SCoRE	10,459	160,337	101,562
TAIYAKU	110,909	1,905,586	1,399,650
TATOEBA	208,013	2,080,831	1,601,860
TED	518,233	4,657,169	3,247,654
WIKIPEDIA	443,849	9,132,894	9,806,199
合計	2,459,913	38,934,507	34,646,840

[1] コーパス・デザインから構築，版権の取得までにかなりのハードルがあるため，単言語コーパスと異なりパラレルコーパスの開発や普及，その活用研究は 10 年単位ではそれほど進んでいない状況である。よって，既存のパラレルコーパスを一覧検索できるように再設計した『(仮称) パラレルリンク』は，既存資源を有効に活用するという点で意味があろう。

なお，大規模ウェブパラレルコーパス JParaCrawl（1000 万対）（http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/）も収録することを検討したが，ノイズが多いことから今回は含めていない。また，アカデミック分野のパラレルコーパス Asian Scientific Paper Excerpt Corpus（ASPEC）（300 万対）（http://orchid.kuee.kyoto-u.ac.jp/ASPEC/）の収録も検討したが，本検索ツールでは一般公開を目指していることから，研究目的に限定されている当該コーパスも含めていない[2]。一方で，仁科（2020）では映画やドラマの日英・英日字幕コーパスを含める有用性に触れていたことから，ノイズが多いものの今回は JESC を含める方針を採った。また，TATOEBA コーパスの元になった田中コーパスは，学生に翻訳させた対訳文を数年かけて収集した約 15 万対からなるコーパスであるが，会話文が全体の 40%を占めることから含めることにした[3]。

なお，教育目的では SCoRE 用例コーパスを活用することが最適であるが，言語学的な分析に関しては SCoRE を除く他 8 種のパラレルコーパスを用いるべきかもしれない（仁科，2020 参照）。また，コーパス間で比較分析（ジャンル・レジスター分析）を可能にするため，各コーパスサイズが異なることも勘案し，コーパスごとに 100 万語あたりの生起頻度をデフォルトで表示する機能も実装予定である。しかしながら，今回搭載予定の 9 種のコーパスだけでは，検索したい語・句の十分な翻訳例が得られない可能性もあるため，Ver.2.0 以降では，今回搭載を見送った JENAAD（日英新聞記事対応付けデータ)や ASPEC（Asian Scientific Paper Excerpt Corpus），Hiragana Times（日英対訳コーパスデータ)に加えて自作コーパスの追加等も検討し，それらのフォーマットを再整備・統一する予定である[4]。

2.2. 『(仮称）パラレルリンク』(Ver.1.0）のテキスト処理・アノテーション

各パラレルコーパスのフォーマットを統一するためにテキスト処理を施し，英語には品詞情報，日本語には形態素情報を付与した。そして，BlackLab query tool を用いて全文検索のインデックスを作成した[5]。

まず，テクスト処理としてテクストのクリーニング，エンコーディングの統一（UTF8），フォーマットの統一，センテンス ID の付与を施した。以下は，対訳ファイルのサンプル（TED）である。

---

[2] 内部利用はおそらく可能であることから，一般公開用の『パラレルリンク』に加えて，使用者を限定した研究者用の『パラレルリンク PRO』の開発も可能な限り進めたい。特に，後者の『パラレルリンク PRO』では，例えば，本稿の第二著者が中心となって独自に構築中の大規模パラレルコーパスを搭載することで，個々の言語研究や翻訳活動，辞書編纂等での活用が一層期待される。

[3] 正確には 146,784 文が日本語と英語の両方で書かれており，短いセンテンスが大半，英文の長さが平均で 7.72 語，最長で 45 語との報告がある（http://hihan.hatenablog.com/entry/2019/01/20/070254）。また，学生 1 人あたり 300 個の文章を翻訳したことから，翻訳者が多数存在する一方で複数の日本人大学生が翻訳プロジェクトに参加したため誤訳が混ざっている可能性もあり，質が保証できないという欠点がある。

[4] ただし，JENAAD は既に無償配布が終了しているため，使用許諾に費用が発生する（JENAAD の有償ライセンスは非商用で 50 万程度である）。同様に Hiragana Times 日英対訳コーパスデータのアカデミックユースは一般の 40%引きの 150 万程度で契約可能である。

[5] BlackLab query tool については以下を参照 https://inl.github.io/BlackLab/query-tool.html.

次に，品詞情報・形態論情報を付与した。まず，英文に関しては，Stanford POS Tagger（https://nlp.stanford.edu/software/tagger.shtml）を用いて，表層形，レマ，品詞など品詞に関する情報を付与した。また，日本語に関しては形態素解析器 Janome（https://mocobeta.github.io/janome/）を使用し，表層形，語彙素，品詞に関する形態論情報を付与した。

■対訳ファイルサンプル（TED）

```
TED  00001  0000000001  I'm going to talk to you tonight　今晩 お話しするのは
TED  00001  0000000002  about coming out of the closet　カミングアウトについてです
TED  00001  0000000003  and not in the traditional sense　いわゆる「カミンアウト」
TED  00001  0000000004  not just the gay closet.　ゲイだと打ち明けることではありません
TED  00001  0000000005  I think we all have closets.　誰しも心に壁を作っています
TED  00001  0000000006  Your closet may be telling someone　その後ろに隠れているのは
TED  00001  0000000007  you love her for the first time　誰かに初めて愛の告白をすることや
TED  00001  0000000008  or telling someone that you're pregnant　妊娠したこと
TED  00001  0000000009  or telling someone you have cancer ガンであることを伝えることかもしれません
TED  00001  0000000010  or any of the other hard conversations　他にも私たちが人生で経験する一
```

2.3. 『（仮称）パラレルリンク』（Ver.1.0）の全文検索インデックスの作成

その後，全文検索インデックスを作成した。詳しくは，上記の品詞情報，形態論情報を含む Blacklab Query Tool（https://inl.github.io/BlackLab/query-tool.html）のインポートファイルを作成した。ファイル形式は XML である。

■インポートファイルサンプル（TED 英文）

```
<?xml version="1.0" encoding="UTF-8"?>
<docs>
  <doc corpus="TED" subcorpus="" fid="00001" sid="0000000001" type="en" counterpart="今晩 お話しするのは">
    <s id="TED::00001:0000000001">
      <w p="PRP" l="I">I</w>
      <w p="VBP" l="be">'m</w>
      <w p="VBG" l="go">going</w>
      <w p="TO" l="to">to</w>
      <w p="VB" l="talk">talk</w>
      <w p="TO" l="to">to</w>
      <w p="PRP" l="you">you</w>
      <w p="RB" l="tonight">tonight</w>
    </s>
  </doc>
```

■インポートファイルサンプル（TED 日本語文）

```
<?xml version="1.0" encoding="UTF-8"?>
<docs>
```

```
  <doc corpus="TED" subcorpus="" fid="00001" sid="0000000001" type="ja" counterpart="l&#x27;m
going to talk to you tonight">
    <s id="TED::00001:0000000001" corpus="TED" type="ja">
      <w p="名詞,副詞可能,*,*" l="今晩">今晩</w>
      <pu> </pu>
      <w p="名詞,サ変接続,*,*" l="お話し">お話し</w>
      <w p="動詞,自立,*,*" l="する">する</w>
      <w p="名詞,非自立,一般,*" l="の">の</w>
      <w p="助詞,係助詞,*,*" l="は">は</w>
    </s>
  </doc>
```

2.4. ファイル整理

処理後のテクストファイルを大きく分けて 3 種のフォルダ（formatted; annotated; blacklab）ごとに整理した。formatted フォルダには，パラレルコーパスの種類ごとに 9 種類のサブフォルダ（JESC; Law; OpenSource; Reuters; SCoRE; Taiyaku; Tatoeba; TED; Wikipedia）が用意されており，エンコーディングは UTF-8 で統一している。annotated フォルダには, formatted フォルダにある統一フォーマットのコーパスデータにアノテーション情報を付与したファイルを収納している。ファイルのフォーマットは XML ファイルで，BlackLab Query Tool のインポートファイルとなる。各コーパスについて，英文と和文の 2 種類の XML ファイルが用意されている。blacklab フォルダには，BlackLab Query Tool のインデックスファイルが収納されている。検索ツールのバックエンドの役割を果たす。

3. 今後の展望

本発表では，過去から現在までの日英・英日パラレルコーパスや検索ツール，それらを活用した研究の変遷を振り返り，これからの展望を述べた。そして，本研究プロジェクトで開発中の網羅型日英・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）のプロトタイプの開発状況を報告した。本ツール(Ver.1.0)の開発が成功すれば，まずは, 現在までに無償公開されている 9 種(予定)の日英・英日パラレルコーパスをオンライン上で串刺し検索することができ，検索語に関するオーセンティックな翻訳例や精緻な翻訳プロファイルを獲得することが可能となる。

なお，本ツール（最終的には Ver.3.0）の完成には約 10 年程度を要することを見込んでおり，理想としては現在公開されている全ての日英・英日パラレルコーパスの搭載を目指す。また，自作コーパスの搭載も検討したい。そして，レキシカルプロファイラーとコンコーダンサーの両検索機能も実装したい。なお，赤瀬川・プラシャント・今井 (2014, p.41)によれば，レキシカルプロファイリングとは，「あらかじめ設定された検索式に基づいて，コーパスから様々なタイプのコロケーションの情報を抽出した結果を，文法パターンごとに整理してユーザに提示するコーパス検索手法」とし，「特定の語彙の文法的振る舞いやコロケー

ションをマクロ的視点から調査できる点に大きな特長がある」としている。参考までに，今後の開発スケジュールについては，表 2 を参照されたい。

表 2 『(仮称) パラレルリンク』(Ver.3.0) までの開発スケジュール (予定)

	収録コーパス	検索方向	検索ツール	検索機能	音声機能
Ver.1.0 *2023 年頃に開発予定	9 種	日→英 (レキシカルプロファイラー)	レキシカルプロファイラー	文法パタン検索，共起語検索など	SCoRE
Ver.2.0 *2027 年頃に開発予定	9 種＋α *有償ライセンスコーパス含む	日→英 (レキシカルプロファイラー)，英→日 (レキシカルプロファイラー)	レキシカルプロファイラー	文法パタン検索，共起語検索，ParaConc の Hot Words 機能など	SCoRE，他のコーパスには Natural Reader を活用するなど方法を模索中
Ver. 3.0 *2031 年頃に開発予定	9 種＋α *有償ライセンスコーパス+オリジナルコーパス含む	日→英 (レキシカルプロファイラー)，英→日 (レキシカルプロファイラー)，日→英 (コンコーダンサー)，英→日 (コンコーダンサー)	レキシカルプロファイラー，コンコーダンサー	文法パタン検索，共起語検索，ParaConc の Hot Words 機能，Dual KWIC，統計解析など	SCoRE，他のコーパスには Natural Reader を活用するなど方法を模索中

最後に，一般ユーザーのみならず，辞書編纂や翻訳・通訳実践 (研究)，対照言語学，言語教育など多岐にわたる分野で変則的な検索にも対応した翻訳コーパス集合体のオンライン検索ツールの開発を目指すべく，各ヴァージョンの完成後には，それぞれの特色や機能，活用事例に関して稿を改め報告したい。

謝辞

本研究は JSPS 科研費 20K00692 の助成を受けたものである。ここに，『(仮称) パラレルリンク』の開発にご協力頂いた第二著者の Lago 言語研究所の赤瀬川史朗代表，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)，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)，ならびに関係者の皆様に感謝の意を示す。

引用文献

赤瀬川史朗・プラシャント・パルデシ・今井新悟 (2014)「NINJAL-LWP の類義語比較機能」『第 6 回コーパス日本語学ワークショップ予稿集』, 41-50.

仁科恭徳 (2020)「日英パラレルコーパス WikipediaKyoto-LWP を用いた和英辞典の記述改善案について－「X を固める」の場合－」『英語コーパス研究』, 27, 1-21.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

多様な指標を組み込んだトピックモデル可視化ツールの開発とテクスト分析への応用

黒田　絢香（大阪大学 大学院生）
kuroda22a@gmail.com

Visualization of Topic Models Using Multiple Measures: LDA for Text Analysis

KURODA Ayaka（Osaka University, Graduate Student）

Abstract

Textual analysis using Latent Dirichlet Allocation(LDA), a generative probabilistic topic model, is recently applied in various fields including literary studies; however, it also requires much effort to mine meaningful topics from huge output data. This study, therefore, aims to develop a simple but interactive visualization tool. We also attempt to reflect various diagnostic measures, such as 'document entropy', 'coherence', and 'effective number of words', in the displayed graph to make exploratory analysis much more easier.

Keywords

トピックモデル, 文学作品分析, Diagnostics, ビジュアライゼーション

1. はじめに

1.1 研究の背景

近年，機械学習アルゴリズムの一つであるトピックモデル，特に Latent Dirichlet Allocation（LDA）（Blei et al., 2003）をコーパス分析に応用する研究が数多く見られる（Jockers and Mimno, 2013; 田畑, 2017; Kuroda, 2018）。合わせて，データの前処理方法や適切なトピック数の設定など，手法の最適化についても議論がなされているが（Lau et al., 2014; Sbalchiero and Eder, 2020），分析対象データの規模や種類，研究目的によって最適解は異なると結論付けられることが多い。そのため，分析の際には少しずつ設定を変化させて出力を適宜確認し，各々の目的に応じて調整する必要があると考えられる。

一方で，トピックがどのような語で構成されているか，各トピックがどのように分布しているのかを示す様々な可視化手法の提案や，それらを容易に描画するためのライブラリの開発も活発に行われている。中でも，pyLDAvis（https://github.com/bmabey/pyLDAvis）は最も多く利用されている可視化ライブラリの一つであるが，トピック数は 10 から 50 程度を想定していること，各トピックがどの文書に出現しているかを図示できないことなど，

いくつかの制限も存在する。

1.2 研究の目的

本研究の目的は，文学作品コーパスのように大規模かつ多様な話題が含まれるデータセットを対象としたトピックモデル可視化ツールを開発することで，試行錯誤を要する分析をサポートし，文学作品の量的研究に寄与することである。ツールの開発には，グラフ作成ライブラリである Plotly，及び Web アプリケーションフレームワークの Dash を用いた。これらを用いてインタラクティブなグラフを出力することで，静的なグラフに比べてより探索的なアプローチを取りやすくすることが今回提案するツールの狙いである。

2. 手法

2.1 データと前処理

本研究ではサンプルとして, Arthur Conan Doyle の小説作品を集めたコーパスを用いた。推理小説や歴史小説，ノンフィクションなどを含む 5 つのジャンル，29 作品から構成される約 220 万語規模のデータセットである。

分析に先立ち，TreeTagger を用いて普通名詞，形容詞，一般動詞のみを取り出した。また今回の実験では，全てのファイルを 1,000 語ごとに切り分けた。

2.2 モデリング

LDA の実装には NLP ツールキットの MALLET（http://mallet.cs.umass.edu/）を用いた。前述の通り，抽出するトピック数はユーザが任意に設定できるため，50，100，150，200 と段階的に数値を変化させて実験を行った。現段階ではそれぞれ個別にグラフを生成しているが，最終的にはアプリケーション上でトピック数を設定することで複数の出力を切り替えられる機能を追加する予定である。

3. 出力データと評価指標

3.1 トピック分布と単語分布

MALLET を利用してトピックモデルを生成した結果，各トピックを構成する単語のリスト，そして文書ごとの各トピック出現確率分布が得られる。これら 2 つの分布を分析することで，各文書において特徴的に用いられているトピック傾向を発見することが，LDA に基づくコーパス分析の基本的な流れとなる。

3.2 diagnostics

MALLET では前述の 2 つに加え，トピックの質を測る diagnostics と呼ばれる XML ファイルを出力することができる。これには各トピックに割り当てられた単語数を示す tokens やトピックを構成する単語の一貫性を評価する coherence，平均単語長を表す word length など 12 の統計指標が含まれており, 膨大なデータから分析目的と合致する有意義なトピックを探し出す，あるいはあまりに周縁的で解釈困難なトピックを除外する際に大いに役立つ。

これらの統計指標についてそれぞれの特徴や関係性を把握するため, 100 トピックを抽出した例から散布図行列を作成した（図 1)。例えば tokens に注目すると，document entropy や corpus distribution と相関関係があり，どれもトピックの generality を示す指標だということが読み取れる。

アプリケーション上ではカーソルを載せるとトピック ID や主要構成単語が確認できるため, 外れ値のような特徴を持つトピックを特定することもできる。例えば，allocation ratio, allocation count は大半のトピックが 0 を示しており，0 から離れるほどごく一部の文書にしか出現しない局所的なトピックであることを表す。

抽出トピック数が多い場合，モデルの全体像を示すグラフはより複雑なものになってしまう傾向がある。diagnostics データから得られる一部の指標を組み込むことで，大量のトピックから目的のものを選び出す探索過程をより短縮できると考えている。

図 1　Diagnostics 指標の相関を示す散布図行列（トピック数 100 の場合）

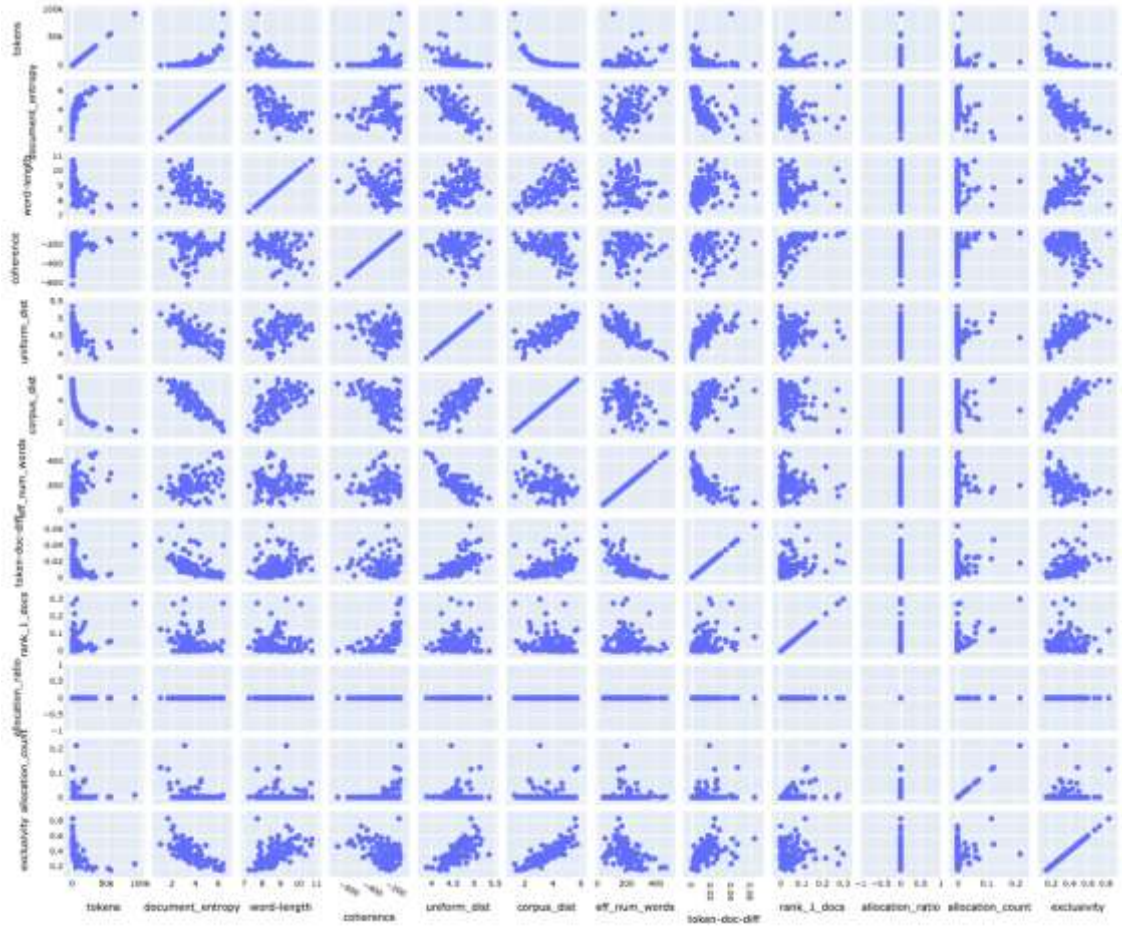
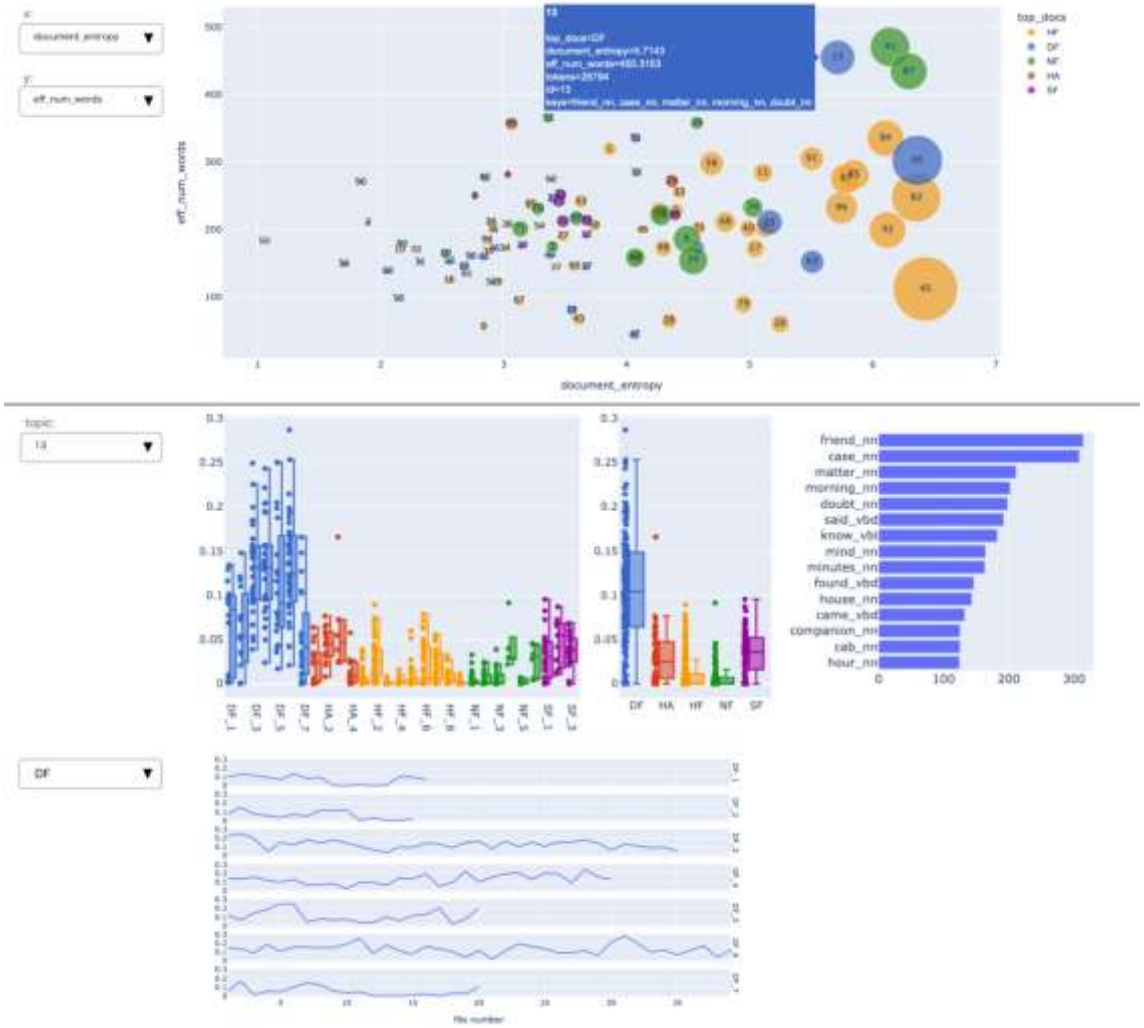


図 2　開発中トピックモデル可視化ツールの概観

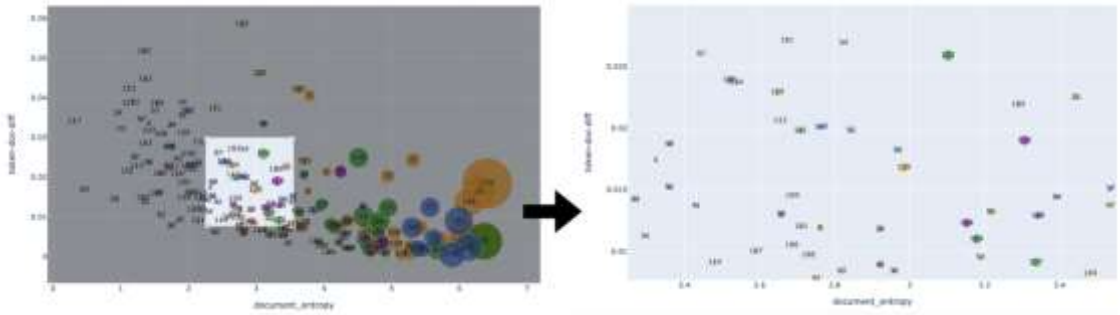


4. トピックモデル可視化ツール

以上を踏まえ，前述の評価指標を組み込んだツールの概観を図 2 に示す。ダッシュボードの上部，overview 部分では任意に選択した 2 つの指標に基づいてモデルの全体像を可視化し，下部では選択したトピックの詳細が確認できる。より詳細なトピック分布を確認するため，最下部では作品ごとのトピック推移を表示した。全ての作品についてグラフを描画するとプロットが大きくなりすぎてしまうため，ジャンルを選択するドロップダウンメニューを追加している。

抽出するトピック数が 100 を超える場合，上部の overview グラフに表示される項目が非常に多く，煩雑なものになってしまう。そこで，カーソルを用いて選択することでグラフの一部を拡大できる仕様とした(図 3)。これにより，トピック数が多い場合でも目的とするトピックを発見しやすくなる。

図 3 overview の拡大機能イメージ（トピック数 200 の場合）



5. 今後の課題

5.1 フィルタ機能

3.2 節で言及したように，allocation ratio や allocation count を用いることで，文書群の特徴をあまり反映しない局所的なトピックを判別することができる。一方で，tokens の値が非常に大きなトピックは，ほとんどの文書に登場する漠然としたトピックであることが多い。今後，それらに合致するトピックを overview から取り除くオプションを追加することで，よりシンプルなプロットを描画できるようにしたい。

また，トピックリストから表示したいトピックのみを選択する機能，特定の文書群に出現するトピックのみを選択する機能も検討している。

5.2 類似トピック表示

MALLET の diagnostics には，トピック間の類似性を評価する指標は組み込まれていない。一方で pyLDAvis では，多次元尺度法（MDS）に基づいてトピックを配置することで，トピック同士の類似性を距離に変換し可視化している。今後は類似性を評価する指標を組み込むことで，overview 部分のグラフ表示を切り替えたり，選択したトピックに類似したトピックをレコメンドする機能を追加したい。

引用文献

Blei, M., Ng, A. and Jordan, M. （2003）. Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Blei, M. （2012）. Probabilistic Topic Models. *Communications of the ACM, 55*（4）, 77-84.

Jockers, M. and Mimno, D. （2013）. Significant themes in 19th-century literature. *Poetics, 41*, 750-769.

Kuroda, A. （2018）. *Topic Representation across Texts and Genres: Finding Key-words through Topic Models*. （Unpublished master's thesis）. Osaka University.

Lau, J., Newman, D. and Baldwin, T. （2014） Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* 530-539.
Sbalchiero, S., Eder, M. （2020）. Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Qual Quant, 54*, 1095-1108.
田畑智司 （2017） 「FLOB コーパスの意味構造: 確率論的トピックモデルによる言語使用域の特徴付け」『統計数理研究所共同研究リポート』*386*, 1--17.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

A Learner Corpus-Based Study of L1 Effects on L2 English Auxiliary Verb Use —The Case of *Will*—

NEWBERY-PAYTON　Laurence
(Tokyo University of Foreign Studies)

Abstract

This study analyzes use of the modal auxiliary *will* in writing by L1 Chinese and Japanese learners of English using data from the International Corpus Network of Asian Learners of English. Lower proficiency L1 Chinese and L1 Japanese learners overuse and underuse *will* respectively. Chinese learners' higher frequency of use can be partially attributed to their use of *will* to express non-future meanings, analogous to functions of the Chinese modal auxiliary *hui*. Japanese learners, who lack L1 functional equivalents to *will*, exhibit underuse as well as omission in obligatory contexts. However, these trends are restricted to one of two essay tasks, suggesting task-related factors.

Keywords

learner corpus, auxiliary verb, modal verb, L1 influence

1. Introduction

This paper analyzes use of the modal auxiliary *will* by Japanese learners of English (JLE) and Chinese learners of English (CLE), using data from the International Corpus Network of Asian Learners of English (ICNALE). Analysis reveals significant differences between JLE, CLE and native speakers (NS) and suggests that the presence or absence in L1 of functional equivalents to *will* may be one cause of differing trends of use.

2. Literature Review

Nakayama (2020) reports that JLE overuse *can*, *should* and *must*, but underuse *will* and *would* in ICNALE's written component. He suggests this reflects the greater difficulty of epistemic modality markers, but does not consider learners' proficiency levels. Nakayama (2021) finds that JLE at A2 and B1 levels underuse *could*, *might*, *would* and *will* in spoken language, and use modal verbs to express deontic modality more frequently than epistemic modality, contrasting with native speakers. However, the analysis combines multiple modal verbs so the findings are difficult to interpret.

Xiao (2017) compares data from learner and native corpora and reports that CLE

overuse *must*, *should*, *will* and *can*, but underuse *would*, *might* and *could* in their writing. Xiao's analytical framework, which groups *will* with other "middle-value" modals, *would* and *shall*, cannot adequately explain the high frequency of use of *will*.

Yang (2018) reports that modal verbs appear more frequently in learners' academic writing than in published academic papers, and that learners overuse *can*, *will*, *could* and *would*. Yang (2018, p. 127) suggests that one-to-one translations of modal verbs in course books may cause pragmatically inappropriate uses of *should* by CLE. The current paper considers the possibility of a similar process occurring in the use of *will* by CLE.

Newbery-Payton and Mochizuki (2020) explore this hypothesis in their analysis of L1 to English translations by CLE and JLE. Errors of omission of *will* appear exclusively in JLE data, while errors by CLE are characterized by inappropriate use of *will* in habitual senses. Newbery-Payton and Mochizuki show that while Japanese lacks equivalent auxiliary verbs, potentially contributing to errors of omission, the Chinese auxiliary verb *hui* shares similarities with *will*, so CLE may equate the two.

Tsai (2015) distinguishes 5 uses of *hui* as a modal verb: ability, future, epistemic, dispositional and generic. While future and epistemic uses correspond to uses of *will*, dispositional (1) and generic (2) modals are less typically expressed using *will*. The two are referred to below as "non-future" uses of *will*. If CLE use *will* analogously to *hui*, we expect increased or even infelicitous use of "non-future" *will* by CLE, but not by JLE.

(1) Waijiaoguan changchang hui lai zheli. [dispositional modal]
diplomat often tend.to come here
'Diplomats often tend to come here.'

(2) Shui hui wang dichu liu. [generic modal]
water HUI towards low.land flow
'Water flows to lower places.' (Tsai, 2015, p. 278)

3. Research Design

3.1 Aim and Research Questions

The current paper develops the analysis of Newbery-Payton & Mochizuki (2020) by a) using a larger data set; b) conducting statistical analyses; c) using a different task format; and d) including comparison of learners at different proficiency levels as well as NS. The research questions are as follows:

RQ1: To what extent do CLE and JLE differ in their use of the modal auxiliary verb *will*?

RQ2: To what extent does the use of *will* by CLE and JLE become more native-like with increasing proficiency?

RQ3: To what extent can the use of *will* by CLE be explained by reference to L1 forms?

3.2 Data and Method

Data is sourced from the Written Essays module of ICNALE (Ishikawa, 2013), allowing consideration of L1-, proficiency- and topic-related factors in the analysis. The two essay topics, "part time job" and "smoking", are abbreviated as "PTJ" and "SMK" below. 48 essays on each topic were randomly selected from the data sets for JLE and CLE at A2, B1-1 and B1-2 level. B2 learners were excluded from the analysis due to data size limitations. Data was processed using AntConc. Table 1 gives a summary of the data.

Table 1 *Data Summary*

	PTJ				SMK		
	A2	B1-1	B1-2		A2	B1-1	B1-2
CHN	10923	11972	12287	CHN	10713	11203	11438
JPN	10933	10540	11057	JPN	10423	10349	10850
NS		10774		NS		10626	
				Total Files: 672 / Total Words: 154,088			

4. Results and Discussion

4.1 Quantitative Analysis

Figures 1 and 2 show adjusted frequency for JLE and CLE on the PTJ and SMK tasks respectively. Black dotted lines show the performance of NS on each task. CLE display more frequent use of *will* than JLE on both tasks, but the two figures otherwise show different characteristics. There is greater variation in the SMK task and an apparent proficiency effect. In other words, learners approach native-like frequency of use as proficiency increases. In contrast, there is relatively little variation between learners and NS in the PTJ task, and little apparent effect of proficiency affecting frequency.

A Kruskal-Wallis test (df=6, χ^2 =57.334230, p= 1.563306e-10) revealed that the groups were not homogenous on the SMK task. Results of post-hoc tests (Dunn method, adjusted with Holm FWER for multiple comparisons) are reported in Table 2. A Kruskal-Wallis test on the PTJ data found no significant difference between groups.

Figure 1 *Adjusted frequency (per 10,000 words) of will in "Part Time Job" task*

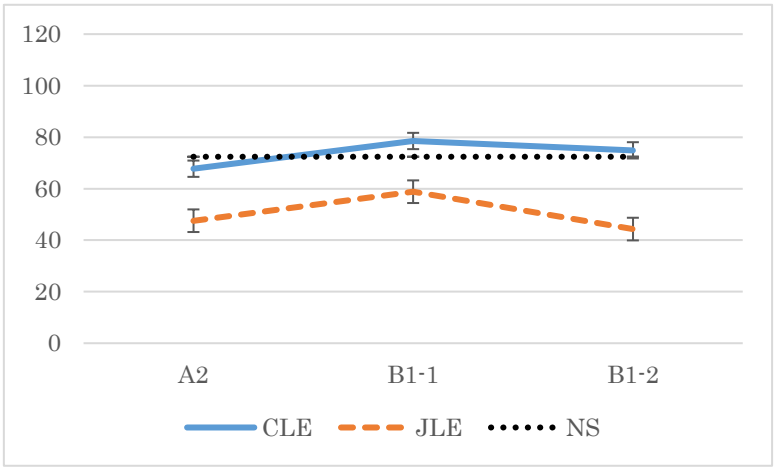


Figure 2 *Adjusted frequency (per 10,000 words) of will in "Smoking" task*

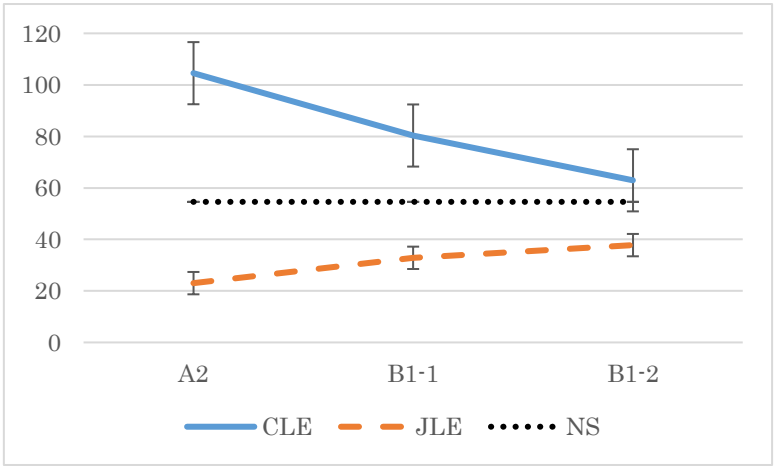


Table 2 *Dunn adjusted p-values for pairwise comparisons (SMK)*

	CLE_A2	CLE_B1-1	CLE_B1-2	JLE_A2	JLE_B1-1	JLE_B1-2
CLE_B1-1	2.89E-01					
CLE_B1-2	**3.52E-02***	1				
JLE_A2	**5.93E-09***	**0.000535***	0.014145			
JLE_B1-1	**5.98E-07***	**0.011524***	0.120829	1		
JLE_B1-2	**1.10E-05***	0.051651	0.364192	1	1	
NS	**1.86E-02***	1	1	**0.027439***	0.207126	0.537358

The significant differences (shown in bold) can be summarized as follows. CLE at A2 level use *will* significantly more frequently than almost all other groups. CLE at B1-1 level also show significantly higher frequency of use than A2 and B1-1 level JLE.

Comparing learners to NS, A2 level learners show significantly higher (CLE) or lower (JLE) frequency than NS. No significant differences were found between learners at B1-2 level or between these learners and native speakers. In other words, non-nativelike frequency of use was limited to lower proficiency levels. This provides answers to RQ1 and RQ2: CLE and JLE differ significantly at A2 and partially at B1-1 level; from B1-1 level onwards, the frequency of use by learners becomes more native-like.

4.2 Qualitative Analysis

RQ3 asked the extent to which the "non-future" uses of the Chinese auxiliary *hui* might influence CLE use of *will*. Instances of *will* were considered "non-future" if they expressed current states of affairs and could be replaced with present tense forms with minimal change of meaning, as in (3). Contrary to expectations, A2 level JLE showed similar uses (4). However, as proficiency increases, non-future uses of *will* become largely limited to CLE (Table 3), providing partial confirmation of the prediction for RQ 3.

(3) There are many people who like smoking, even in the public places they **will** take a cigarette in hand. W_CHN_SMK0_289_A2

(4) Especially, in the restaurant, many people **will** enjoy eating and talking with friends or families. W_JPN_SMK0_344_A2_0

Table 3 *Frequency and proportion of non-future uses of will*

	A2	A2 (%)	B1-1	B1-1 (%)	B1-2	B1-2 (%)
CLE	20	0.178571	18	0.2	11	0.152778
JLE	5	0.208333	3	0.085714	1	0.02439
NS	1	0.017241	1	0.017241	1	0.017241

Finally, omission of *will* in conditional clauses – where Chinese uses *hui*, but Japanese requires no overt morphological form – are largely limited to JLE. The presence or absence of a (partially) equivalent morphological form in L1 may lead to increased or reduced salience of the L2 form for JLE and CLE respectively.

Table 3 *Omission of will in obligatory contexts (conditional clauses)*

	A2	A2 (%)	B1-1	B1-1 (%)	B1-2	B1-2 (%)
CLE	3	0.037037	3	0.058824	2	0.05
JLE	10	0.138889	12	0.173913	13	0.168831

4.3 Task-related effects.

As already stated, differences in the PTJ task were not statistically significant. While neither essay topic discourages use of *will*, SMK is arguably more conducive to writing about hypothetical future events. Adjusted frequencies are higher in 5/6 learner groups (the exception is B1-2 CLE), but for NS, frequency of *will* is actually higher on the PTJ task. More thorough analysis of task-related effects will be required in future.

5. Conclusion

This study revealed significant differences in the use of *will* by CLE and JLE at lower proficiency levels, and has suggested one cause. Analysis was limited to L2 writing, so spoken data from ICNALE should also be analyzed in future. Online processing demands during speech may cause higher rates of omission, particularly at lower proficiency levels, though how this relates to L1- and task-related effects remains to be seen.

Bibliography

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. School of Languages & Communication, Kobe University. *Learner Corpus Studies in Asia and the World*, *1*, 91-118.

Nakayama, S. (2020). Contrastive interlanguage analysis of modal auxiliary verb usage by Japanese learners of English in argumentative essays. *The IAFOR International Conference on Education – Hawaii 2020 Official Conference Proceedings*.

Nakayama, S. (2021). *Modal auxiliary verbs in Japanese EFL learners' conversation: A corpus-based study. Asia Pacific Journal of Corpus Research*, *2*(1), 23-34.

Newbery-Payton, L., & Mochizuki, K. (2020). L1 influence on use of tense/aspect by Chinese and Japanese learners of English. *Learner Corpus Studies in Asia and the World*, *4*, 67-93.

Tsai, W. (2015). On the topography of Chinese modals. In U. Shlonsky, (Ed.), *Beyond functional sequence* (pp.275-294). Oxford University Press.

Xiao, Y. (2017). Chinese ELF learners' acquisition of modal verbs: A corpus-based study. *International Journal of English Linguistics*, *7*(6), 164-170.

Yang, X. (2018). A corpus-based study of modal verbs in Chinese learners' academic writing. *English Language Teaching*, *11*(2), 122-130.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 2021

授業コーパス構築のための自動タグ付けツール "*Classroom Corpus Tagger*"の開発

大橋 由紀子(ヤマザキ動物看護大学)
y_watanabe@yamazaki.ac.jp
片桐 徳昭(北海道教育大学)
katagiri.noriaki@a.hokkyodai.ac.jp
押切 孝雄(戸板女子短期大学)
oshikiri@toita.ac.jp

Developing Classroom Corpus Tagger: A Spoken Language Tagger to Compile Classroom Corpora

OHASHI Yukiko (Yamazaki University of Animal Health Technology)
KATAGIRI Noriaki (Hokkaido University of Education)
OSHIKIRI Takao (Toita Women's College)

Abstract

This study introduces the Classroom Corpus Tagger (CCT), which automatically generates tags for transcribed utterances by lines. The construction of a classroom corpus requires tagging of speakers, languages, and activities according to the annotation design (e.g., Ohashi & Katagiri, 2016). The CCT uses JavaScript and automatically determines the language and assigns language (Japanese or English) tags. Multiple types of speaker tags can be set as desired in the CCT, which will also be attached automatically to the transcribed utterances. As a result of an experiment examining the validity of the CCT tagging, the errors seen in manual construction were not observed when the CCT was used, showing that CCT was more accurate and less burdensome.

Keywords

授業コーパス, 授業分析, reflective practice

1. はじめに

英語授業コーパスの構築は、授業を振り返り、改善へと導く reflective practice の材料となる。Walsh(2013)によると、授業内の発話を書き起こし、インタラクションを観察することで大きな気づきや新たな知見が得られるとされている。更に、自身の授業のコーパスを構

築し、他の授業と比較することで教師教育の材料ともなり得る。授業コーパス構築には利点がある一方で、コーパス構築に要する時間や労力の問題は解決できていない。手動でのコーパス構築はタグ付与でのミスも起こりやすい。そこで本研究では、授業コーパス構築を容易にするため、発話者と使用言語に対する自動タグ付与ツール開発に取り組んだ。

2. 先行研究

授業コーパスの構築には、IC レコーダー等の機材で音声を録音し、書き起こしを行う。コーパスから必要な発話や分析に要するインタラクションを抽出するために、授業コーパスは XML（Extensible Markup Language）インスタンス(以下 XML と表記)で階層構造にて構築すると便利である。XML を使うことで、使用者はタグを研究目的に応じて定義し、多様な情報を「意味」と「内容」に分けて記述することができる。例として、Katagiri & Kawai (2016)は XSLT（XSL Transformations）を用いた談話構造可視化のためのスキーマを提案し、XSLT から授業内での会話を抽出する例や、授業コーパスを会話分析に利用する方法を紹介した。

作成した授業コーパスは XSLT や Perl スクリプトなどで目的データの抽出を行う。日本語発話の書き起こしの際、漢字仮名交じりとなると文字化けの問題などが出現し、ファイルそのものの読み取りの信頼性が欠けることがあり、Perl スクリプトでデータを抽出する際の障害となりうる。そのため、自作コーパスで使用するエンコード形式の事前設定として、Unicode 系の使用が推奨された（片桐・大橋 2016）。

現在までに日本での英語授業を録画、書き起こした内容から構築した授業コーパスの例として、Ohashi (2015)、Ohashi & Katagiri (2016)が挙げられる。Ohashi (2015)では発話の種類に合わせて 24 種類の異なるタグを付与した。活動内容、教員のインプット、学習者のアウトプット、機材の使用等に関する詳細なタグが付与されている。タグ付与は手動であり、複数人数でタグを付与した。複数間でコーパスを構築する場合には作成者相互の理解の相違が生じた場合を想定し、カッパ係数など、評価者間信頼係数にて妥当性を示す等の対策を要した。Ohashi & Katagiri (2016)では教師の明示的説明が学習者の理解にどの程度影響を与えるのかを探るため、教師の発話に<explicit></explicit>と分類するタグを付与し、発話を抽出後、数値化することで分析に利用した。図 1 は Ohashi & Katagiri (2016)より抽出した発話例である。

図 1 授業コーパスの発話例(Ohashi & Katagiri, 2016)

```
    <st><eng>Please stay here.</eng></st>
    <st><eng>Please stay here.</eng></st>
    <hrt><j>って言ってるね。はい。</j></hrt>
  <explicit id="1">
    <hrt><j>これどういう意味なのかってことですね。はい。</j>
          <j>だからここに居てって意味ですね。</j></hrt>
    </explicit>
```

授業コーパスは語彙レベル調査に使用することも可能である。Ohashi & Katagiri (2020) では、日本での外国語授業における語彙使用に焦点をあて、小学校英語授業コーパスを構築後、教師の発話のみを抽出し、CEFR-J wordlist で示される各語彙レベルに相当する語彙の使用率を調査した。コーパスに含まれた 4 クラスの調査で、小学校で使用を推奨されている A1 レベル語彙の使用率は平均 11.8%と低く、授業内での語彙使用を増やすことが提案された。

このように、授業コーパスを構築し、授業内での発話、活動等のデータを数量化、可視化することで分析の変数として利用することが可能となる。雑談などの質的データであっても、学習者の理解に繋がる発話を探るためのソースとして利用できるため、研究のために活用できる一方で、手動でのタグ付与は大量の時間を要し、労力も大きいため、容易に構築することは叶わない。

タグ付与の手作業によるエラーを最小限にするため、正規表現を利用した置換機能によるタグ付けのための表現リストを公開し(大橋ほか, 2016)、動物看護英語教育のための動物病院英文カルテコーパスを構築した。正規表現を利用してタグ付けを半自動化することでコーパス構築の時間は大幅に縮小されたものの、作成するコーパスの種類に応じて使用する正規表現の書き換えを要することや、特殊な発話に対するタグ付与にはやはり手動で付与する必要がある等の問題点は残された。そこで本研究では、授業コーパス構築を更に容易にすべく、話者タグと言語タグを自動的に認識し、発話の質に合わせて半自動でタグが付与されるツールに取り組んだ。

3. Classroom Corpus Tagger

3.1 Classroom Corpus Tagger 利用環境

授業コーパス構築のための自動タグ付与ツールは JavaScript を使用して開発し、Classroom Corpus Tagger（以後、CCT）と名付けた。Windows/Mac 両方に対応している。CCT はダウンロード版であり、ブラウザで動作する（Google Chrome、Firefox、Edge に対応。IE は非推奨)。

3.2 CCT 利用方法

CCT 利用方法を以下に示す。図 2 は実際の画面である。

1. 話者タグ入力ボックス

話者タグは初期値として「hrt st sts」がセットされている（hrt = homeroom teacher, st = student, sts = students の意)。話者タグは任意に設定が可能である。例えば学習者を示すタグとして「st」ではなく、「st1」「st2」と区別したタグに設定できる。話者タグは、空白を含めて 300 文字まで入り、300 文字を越えなければタグの数の制限はない。例えばタグが 5 文字なら 50 タグまで、2 文字なら 100 タグまで設定が可能となる。

2. 書き起こし文入力ボックス

「2」に書き起こし文を入力する。もしくは書き起こしテキストをコピー&ペーストする。XML 化したい文章は<body>と</body>で挟まれたエリアに入力する。あらかじめ作成した文章を貼り付けるか、文字を直接打ち込んでいく。ただし、右画面の XML の行をクリックしはじめたら、左画面の「書き起こし文入力ボックス」に修正を加えることはできない。クリックすると話者タグがすべてリセットされる。書き起こし文の行は 1000 行以上の入力が可能である。

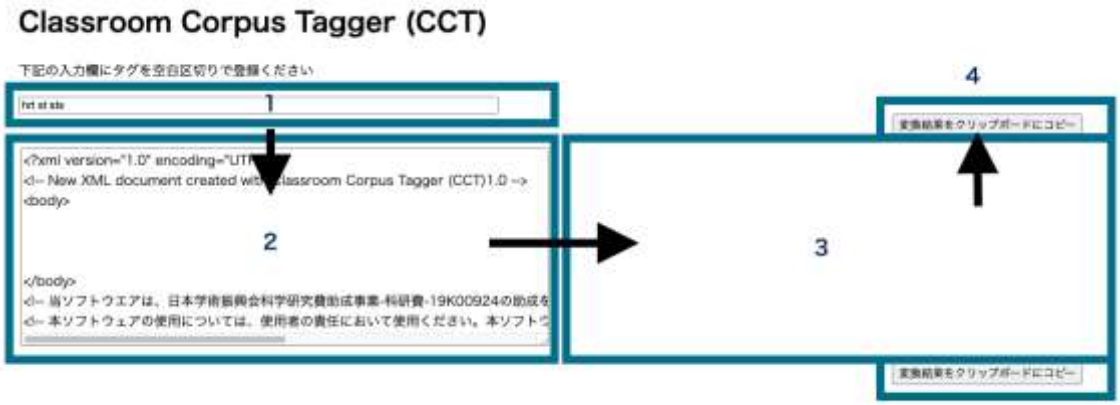
3. XML タグ自動生成

「2」ボックスに文字が入力されると画面右側に自動的にタグ付きの XML インスタンスが生成される。例えば「Hello」と入力すると、<hrt><eng>Hello</eng></hrt>と即座に表示される。話者タグは「1.話者タグ入力ボックス」の先頭の文字列が表示され、変換結果行をクリックすると hrt → st → sts → hrt...（その他設定したタグ）と切り替わる。言語タグについては、日本語の発話には<j></j>、英語の発話には<eng></eng>、日本語と英語が混在する場合は <mix></mix>が自動的に付与される。例えば「こんにちは」と入れると、<hrt><j>こんにちは</j></hrt>という結果が表示される。英語と日本語が混合した文章の場合、「Hello こんにちは」を入力すると、<hrt><mix><eng>Hello</eng><j>こんにちは</j></mix></hrt>と表示される。

4.「変換結果をクリップボードにコピー」

入力作業終了後、「変換結果をクリップボードにコピー」ボタンを押すことで、XML インスタンスをコピーできるため、その後は任意のテキストエディタなどへ貼り付けて保存が可能となる。

図 2 CCT で示される画面



1.話者タグ入力ボックス 2.書き起こし文入力ボックス 3.XML タグ（インスタンス）生成スペース 4.「変換結果をクリップボードにコピー」ボタン

4. CCT の評価

本節では、既存の授業の書き起こし（プレーンテキスト）を用いた CCT の動作確認と結果の評価をする。以下、書き起こしテキスト、手動によるタグ付け結果との比較（4.1）、手動タグ付け作業者による使用感想に基づいた考察（4.2）について順に記載する。評価資料として教育実習生（教員養成系国立大学 3 年生）2 人の研究授業（中学 1 年 1 クラス、2 年 1 クラス）を書き起こしたものを利用した。

4.1 CCT と手動によるタグ付結果

教育実習生の研究授業の 2 クラス分の書き起こしプレーンテキストを用い、話者タグ（教師と生徒）及び使用言語タグ（日本語、英語、日英のミックス）の開始タグと終了タグについて正確さを評価した（表 1）。

表 1 手動タグ付けと CCT での教員発話 token の比較

実習生研究授業番号	発話数		差異
	手動でのタグ付け	CCT でのタグ付け	
1	267	203	64
2	145	117	28

表 1 に手動によるタグ付と CCT によるタグ付の結果を記す。手動と CCT でのタグ付けとして得られた総数に差が生じている。これらの差について目視で確認すると、以下の傾向が見られた。

1) アポストロフィー(') などの非文字表記を本来 1 バイト文字（半角）で打つべきところを、全角（2 バイト）文字で入力した場合、使用言語が日本語(<j></j>)と CCT が判断し、誤りに繋がる。

2) 日本語の固有名詞等を 1 バイト文字（アルファベット）表記した場合は使用言語が英語(<eng></eng>)と CCT が判断するため、固有名詞の表記ミスは誤作動に繋がる。

4.2 考察

本研究では、同じ授業の書き起こしに対して、同一人物による、人力による手動タグ付と CCT によるタグ付とを行い、両者の使用感についてアンケート調査を行った。その結果、CCT の利点として、XML 構造を利用したタグの階層（入れ子構造）による複雑さが軽減されること、それにより、タグ付の時間が短縮されることを確認した。特に使用言語タグのうち、mix タグを付与する労力の軽減と時間短縮の利点が挙げられた。

以上 CCT 利用に関する見解を総合すると、英語の授業コーパスを構築する際に CCT を利用することは、正確さと労力・時間の上で、手動によるタグ付と比較して優位であることが判明した。

5. まとめ

英語授業コーパスを構築する場合、CCT を利用することにより、手動でタグ付をする事と比較し

て正確性と、時間短縮による労力の軽減の点で優位であることを論じた。本節では CCT 利用の注意点と今後の改良点を論じ、本研究のまとめとする。

CCT 利用の留意点として、言語の読み込みの問題が挙げられる。2 バイト文字へは<j>、1 バイト文字へは<eng>を付与している。そのため、他の 2 バイト文字の言語である中国語や韓国語を入力しても<j>を表示する。同様に、1 バイト文字であるかぎり、仮にフランス語やドイツ語を入力しても<eng>と表示される。よって、当面、CCT は日本語と英語専用のツールである点に注意しなければならない。また、本ツールは、エディタではなく、タグ付与を効率化するための Tagger であるため、エディタの機能は現在持ち合わせていない。

CCT の今後のバージョンアップとして、文字バイト数のみに依存しない使用言語タグ付与方法を開発して、日本語-英語以外の言語に対応させる事と、エディタを使って授業を書き起こしする段階からのタグ付与機能実装を試みたい。

謝辞

本研究は JSPS 科研費 19K00924 の助成を受けたものです。

引用文献

Katagiri, N., & Kawai, G.(2016). Designing XML schema for classroom discourse visual representation through XSLT. *Journal of Hokkaido University of Education (Humanities and Social Sciences), 66*(2), 1-16.

Ohashi, Y. (2015). A Corpus-based study on the relationship between the languages used in junior high school classrooms and learners' uptake. *KATE Journal. 29*, 29-42.

Ohashi, Y., & Katagiri, N. (2016). The effects of explicit instructions observed in teacher transcripts and student impression remarks in elementary school. *HELES Journal (16)*, 3-18.

Ohashi, Y., & Katagiri, N. (2020) The Ratios of CEFR-J vocabulary Usage Compared with GSL and AWL in Elementary EFL Classrooms and Suggestions of Vocabulary Items to be Taught. *Asia Pacific Journal of Corpus Research 1*(1), 35-65.

Walsh, S. (2013). *Classroom Discourse and Teacher Development*. Edinburgh.

大橋由紀子・岡勝巖・関谷弘毅・花田道子 (2017)「正規表現を利用した動物病院英文カルテのコーパス化と英語教育への応用」『ヤマザキ学園大学紀要』 第 7 号, 25-40.

片桐徳昭・大橋由紀子（2016）高汎用性教室英語の発話コーパス構築の課題と蓄積の方向性『北海道教育大学紀要（人文科学・社会科学編）』*67*(1). 15-25.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

日本人学習者の英語原因表現使用：ICNALE に基づく量的概観 ―原因表現 34 種の使用実態の解明―

佐々木　恭子（神戸大学 大学院生）
zuomug621@gmail.com

Use of English Causal Expressions by Japanese EFL Learners: A Quantitative Analysis based on ICNALE ―How do Japanese EFL Learners Use 34 Causal Expressions in their Essays?―

SASAKI　Kyoko (Kobe University, Graduate Student)

Abstract

It is pointed out that Japanese learners of English (JLE) overuse 'because' in English essays (Kobayashi, 2009; Sasaki, 2021), but this has not been sufficiently verified with a quantitative analysis of a wide range of causal expressions in English. For this purpose, this study investigates the use of causal expressions by JLE comparing to that by English native speakers (ENS) on ICNALE. The data are from 5 sections: JLE at CEFR A2, B1-1, B1_2, B2 level and ENS students. Focusing on 34 expressions from the findings of Altenberg (1984), Biber et al. (1999) and others, the obtained results are: (1) Significant difference of 'because' frequency by JLE is not observed on the corpus. (2) The way of expressing the cause differs between JLE and ENS. (3) JLE' s use of causal expressions does not show linear development according to increased learners' proficiency.

Keywords

英語原因理由表現, because 過剰使用, 日本人学習者, 英作文

1. はじめに

論理的な文章を書く上で原因と結果の関係を明示化することは必須であるが, その際, 表現多様性にも注意を払う必要がある。日本人学習者の英作文においては, because の過剰使用が広く指摘されているところであるが（小林, 2009;佐々木, 2021）, because 以外の原因表現の使用状況調査, 特に学習者の習熟度段階の違いを考慮に入れた調査は, いまだ十分に行われていない。

そこで本研究では, 動詞 19 種（allow など）, 接続詞 5 種（because, as など）, 名詞 3 種（reason など）, 前置詞句 7 種（due to など）からなる 34 種の原因表現を取り上げ, 学習者の使用実態を詳細に調査する。調査にあたっては, 日本人大学生と英語母語話者による統制的英作文を収集した既存コーパスを使用し, 日本人学習者による原因表現使用の課題を明らかにする。

2. 先行研究

英語原因表現の枠組みは曖昧であり，さまざまな研究でそれぞれに異なる表現が取り上げられていることも多い。例えば，動詞において Quirk et al.（1985）は causative verbs を，続く不定詞部分が結果を表す動詞と定義づけ，cause, drive, get など 9 動詞を挙げている。このことから，主部情報が原因となり，動詞が原因と結果を繋ぐ役割をすることが分かる。Biber et al.（1999）は，causative verbs は人・無生物が事態に新たな状態を生じさせる動詞と定義し，allow, cause, enable など 8 動詞を例示した。Girju（2003）は，因果関係表出動詞として lead to, make, bring 等 60 種の causative verbs を挙げている。動詞以外の因果表現において Altenberg（1984）は因果表現リストを作成し，各言語使用域の分布と使用の決定要因を調査し，A. so, therefore 等の副詞結合（20 種），B. because of 等の前置詞結合（15 種），C. because, as 等の接続表現（11 種），D. that's why 等の句統合結合（45 種）の合計 91 種を挙げている。以上の原因表現の総数は延べ 168 種となる。

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，母語話者と日本人学習者による 34 種の英語原因表現使用を比較し，日本人学習者の原因表現使用の課題を明らかにすることである。学習者調査にあたっては，日本人学習者の多様性をふまえ，Common European Framework of Reference for Languages（CEFR）に基づく A2，B1-1，B1-2，B2+の 4 つの習熟度段階を区別する。これを踏まえ，3 つの研究設問を設定した。

RQ1　母語話者及び日本人学習者群の間で原因表現の総頻度・品詞別頻度に差があるか。
RQ2　母語話者及び日本人学習者群の間で頻度に差がある原因表現はどれか。
RQ3　原因表現において母語話者及び日本人学習者群はどう分類されるか。

3.2 調査対象

まず，Quirk et al.（1985），Biber et al.（1999），Girju（2003）が causative verbs とする動詞 74 種を選んだ（重複を除く）。その後，高校生の英作文指導への応用を考慮し，西出・水本（2009）の単語親密度リストで 5 段階中 4 以上の 19 種に絞った。動詞以外の表現として Altenberg（1984）は副詞（20），前置詞（15），接続表現（11），節表現（45），計 91 種を挙げている。本研究は原因表現のうち，重複と検索困難語等を除き，計 34 種を調査対象とした。

・動詞（19 種）：allow, bring, cause, contribute, create, develop, drive, effect, enable, force, get, give rise to, help, lead to, link, make, produce, relate to, require
・接続詞（5 種）：as, because, for, since, so that
・名詞（3 種）：　cause, explanation, reason
・前置詞表現（7 種）：because of, due to, for reasons, for the sake of, on account of, on the ground that, owing to

3.3 データ

本研究では International Corpus Network of Asian Learners of English（ICNALE）の Written Essay Module（Ishikawa, 2013）に含まれる日本人大学生と英語母語話者大学生（ENS）の英作文データを使用する。このコーパスでは，トピックが「大学生アルバイトの是非」と「レストラン全面禁煙の是非」の 2 種に統制されているが，本研究ではその両方を分析対象とする。1 作文の長さは 200~300 語，執筆時間は 20～40 分，辞書使用は禁止されている。ICNALE では，学習者は TOEIC 等の英語能力テストまたは語彙サイズテストのスコアを基に，CEFR の 4 段階に区分されている。本研究では，この段階別にサブコーパスを作成し調査する。各サブコーパスのサイズは，A2 が 68,529 語，B1-1 が 79,591 語，B1-2 が 22,390 語，B2＋が 8,532 語である。また，ENS サブコーパスでは 45,028 語である。

3.4 手法

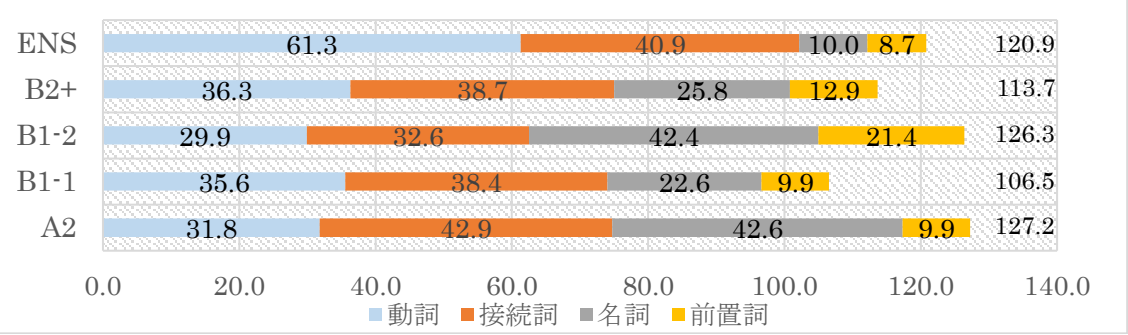
RQ1 では，ENS を含む全 5 群の書き手について，34 表現の総頻度と，動詞・接続詞・名詞・前置詞別頻度を求める。相互比較のため，頻度は 1 万語あたりに調整する。なお，34 表現の出現例の中には原因表現以外も含まれる。そこで，原因表現であると自明な 5 種（because（of）, cause（n, v）, reason ）以外を悉皆調査し，原因表現とされたもののみ計上する。頻度の差の検定には Fisher の正確確率検定を用いる。また，検定の反復のため有意水準を引き下げ，$\alpha = 0.1\%$で判断を行う。RQ2 では 34 表現の各々と ENS 頻度を比較し，学習者が有意に ENS より多く，または少なく使用する表現を特定する。RQ3 では，書き手群を第 1 アイテム（5 カテゴリ），原因表現を第 2 アイテム（34 カテゴリ）とする頻度表に対し対応分析を実施し，得られた第 1・第 2 次元で作成した散布図を質的に解釈する。

4. 結果と考察

4.1 RQ1 総頻度と品詞別頻度

34 種の原因表現に関して総頻度と品詞別頻度を調査したところ，図 1 の結果を得た。

図 1　総頻度・品詞別頻度（1 万語あたり）



総頻度（グラフの右端数字）および品詞別に，書き手群別の頻度を比較し，多いものから順に降順で並べ，その後，隣接間で検定を行ったところ，次の結果を得た。

表 1　書き手群間の順位性と差の有無（頻度順降順）

種別	順位関係と隣接間の差の有無	種別	対母語話者で差があった群
全体	A2≒B1-2≒ENS≒B2≒B1-1	全体	なし
動詞	ENS≒B2≒B1-1≒A2≒B1-2	動詞	B1-1, A2, B1-2（少用）
接続詞	A2≒ENS≒B2≒B1-1≒B1-2	接続詞	なし
名詞	A2≒B1-2≒B2≒B1-1＞ENS	名詞	A2, B1-2, B2, B1-1（多用）
前置詞	B1-2≒B2≒A2≒B1-1≒ENS	前置詞	B1-2

ここでは，日本人学習者内の順位関係と，母語話者との関係の 2 観点から考察を行う。まず前者では，A2～B2＋の 4 段階で使用頻度の増減が一貫するものはなかった。これは，学習者全般の習熟度上昇と，原因表現の発達が必ずしも一致していないことを示唆する。次に後者では，学習者（の一部）が母語話者より多用するのは全ての群での名詞と B1-2 での前置詞，少用では B2＋以外での動詞であった。このことは，学習者は全体的に名詞で原因表出し，母語話者が多用する動詞使用は習熟度が上がると可能になることを表すといえる。

以上より，原因表現の総頻度については母語話者・学習者間，学習者の各習熟度間のいずれにおいても統計的に差がないこと，また，品詞別では母語話者・学習者間の場合に動詞，名詞，前置詞で，学習者の習熟度間では全てにおいて有意な差がないことが確認された。

4.2 RQ2 学習者が多用･少用する表現

前節では，34 種の原因表現を全体として，あるいは品詞別にまとめた場合，母語話者・学習者間，学習者の習熟度間で必ずしも大きな違いがないと示された。続いて 34 種の原因表現を個別に調査すると，母語話者との間で頻度に差のある表現として，以下の結果を得た。

表 2　母語話者頻度と差がある原因表現

差のタイプ	A2	B1-1	B1-2	B2+
多用	reason	reason	reason, because of	reason
少用	allow, help, create, develop, relate to, as	allow, help, create, develop, relate to, so that, due to	allow, help	

まず，日本人が多用する表現に関して 2 点に注目する。1 点目はすべての習熟度群で reason が多用されていることである。

(1) The first <u>reason</u> is that smoking is harmful for human.　(JPN_SMK0_085_A2)
(2) I have two <u>reasons</u> about that.　(JPN_PTJ0_031_B1-2)

こうした表現はアカデミックライティングの定型表現として高校生向け英作文教材などでも例示される（大矢, 2005 など）。学習者はこれらに強く影響され，原因を示す際に reason を多用したものと考えられる。2 点目は，先行研究で日本人学習者による過剰使用が報告されてきた because が今回の調査では多用とならなかったことである。母語話者と学習者が同じ条件で同じ内容について作文を行う統制コーパスを利用したことで，日本人による because の使用が必ずしも過剰ではないことが確認できた。

次に，日本人の少用表現に関し 2 点に注目する。1 点目は母語話者に比べ少用される原因表現の数が A2 から B1-2 にかけて減少し，B2+では該当がなくなることである。これは，習熟度上昇とともに使用できる原因表現が増え，B2＋ではほぼ問題が解消することを示す。加えて，B1-1 と B1-2 間に大きな差が生じることも興味を引く。原因表現の多様性の習得では，A2，B1，B2 の習熟度の境界とは別に，B1 内部で質的変化が生じる可能性が高い。2 点目は，allow と help が B2＋以外で使用が少ないことである。ここで，以下の母語話者の help の用例を見ると，学習者が苦手とされる無生物が主語の構文であることに気づく。

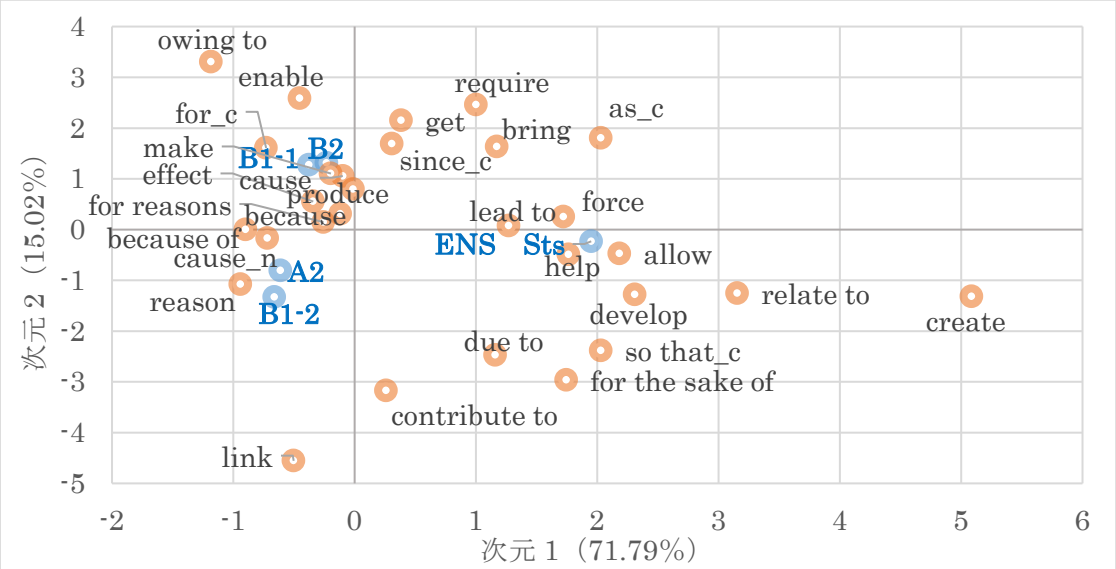
(3) A part-time job can teach you to be responsible, help you make new friends, and help you meet people to network with. （ENS_PTJ0_009_1）

初級学習者は allow を（人が人を）「許す」，help を（人が人を）「助ける」といった原義でのみとらえているため，こうした発想が行いにくい。一方で，上級学習者，さらに母語話者になれば，こうした表現を幅広くうまく使って，原因表現の偏りを解消できるようになる。

4.3 RQ3 母語話者と学習者の分類

対応分析により，以下の結果を得た。第 1 次元・第 2 次元の累積寄与率は 86%を超えており，分散の過半が以下の散布図で集約できているといえる。

図 2 対応分析に基づく散布図



ここでは 4 点に着目する。1 点目は，第 1 次元（横軸）で母語話者が右側に，学習者全群が左側に付置されていることである。これは，個々の原因表現使用で，母語話者と学習者間に根本的差異が存在することを示す。2 点目は，第 2 次元（縦軸）上で，A2 と B1-2 が下部，B1-1 と B2＋が上部に付置されていることである。第 2 次元は習熟度上昇を示す軸と解釈できるが，重要なことは，B1-2 と B1-1 の位置の入れ替りである。原因表現の習得では，B1 内部である種の逆転が起こっている可能性がある。3 点目は，学習者の習熟度上昇

が必ずしも母語話者への近接を意味していないことである。習熟度上昇は図の上部への移動として示されるが，母語話者は 1 軸の右側に区分されているので，現状では母語話者らしい原因表現使用には届かない可能性が強く示される。最後に 4 点目は，各種原因表現がおよそ，ENS を特徴づける第 1・第 4 象限，B1-1 と B2＋を特徴づける第 2 象限，A2 と B1-2 を特徴づける第 3 象限の 3 種に区分されることである。こうしたデータをうまく使えば，各段階の学習者に次位レベルの原因表現を示す教育的介入も可能になるかもしれない。

5. まとめ

調査から，日本人学習者が少用する原因表現に焦点を当て，習熟度別に以下にまとめた。

表 3　日本人学習者の習熟度・品詞別少用原因表現

少用表現のある群	接続詞	動詞	前置詞
(a) A2, B1-1, B1-2		allow, help	
(b) A2, B1-1		develop, create, relate to	
(c) B1-1 のみ			due to, so that
(d) A2 のみ	as		

表 3 の少用傾向(a)～(d)は，教育現場で扱う順序を表す。少用表現は全て教育的援助を要する現象といえるが，その必要性に段階があると考えられる。(a)(b)の表現は複数レベルで使用できず，日本人学習者全体の課題だと考えられるため，優先的に扱うことが重要である。特に日本人学習者の多くが使えない原因表現である allow, help などの動詞で，無生物を主語に置く英語特有の発想への転換の重要性を示し，全体で有意に高頻度の reason の使用についても用例を用いて指導することで，均整の取れた原因表現使用を目指す必要がある。

引用文献

Altenberg, B. (1984). Causal linking in spoken and written English. *Studia Linguistica*, *38*(1), 20-69.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education Ltd.

Girju, R. (2003). Automatic detection of causal relations for question answering. *MultiSumQA '03: Proceedings of the ACL 2003 workshop on multilingual summarization and question answering, 12*, 76-83.

小林雄一郎(2009)「日本人英語学習者の英作文における because の誤用分析」『関東甲信越英語教育学会研究紀要』*23*, 11-21.

西出公之・水本篤(2009)「英単語 8000 語についての親密度測定の試み」『都留文科大学大学院紀要』 *13*, 57-92.

佐々木恭子(2021)「高校生の英作文に見る because 使用:頻度・文中位置の視点から」『統計数理研究所共同研究リポート』*444*, 139-158.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman Group Limited.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

工学系大学院生のための教材開発：日英コーパスの分析 ―自律的な工学英語の学びを支援する新しい 工学論文アブストラクト検索システム ERAP Online の開発―

石川　有香（名古屋工業大学）
ishikawayuka.jp@gmail.com

Developing a New Autonomous Learning System for Graduate Students of Engineering Using the ERAP Corpus

ISHIKAWA Yuka (Nagoya Institute of Technology)

Abstract

This study introduces a Japanese-English parallel corpus, ERAP Corpus, which is compiled from engineering research abstracts. The study analyzes two subcorpora of the ERAP Corpus, the computer science abstracts and the chemical engineering abstracts, aiming to develop a new online autonomous learning system for graduate students of engineering. The results suggest that keyphrases rather than keywords extracted from Japanese abstracts will help learners understand the overall structure of the abstract.

Keywords

ESP, JSP, 日英パラレルコーパス, 工学系論文, 教材開発

1. はじめに： ERAP コーパスの概要

筆者の研究室では，工学学生（主として大学院生）の英語論文執筆をサポートするデータやツールの開発を行っている。これまでに，大型の論文コーパスの分析結果をふまえて工学英語語彙リスト（石川，2017）を開発・公開したほか，現在は，工学論文のアブストラクトを収集した Engineering Research Abstract Parallel Corpus（ERAP）を開発中である。

海外の英語論文や，そのアブストラクト部分をアーカイブしたコーパスはすでに多数存在するわけだが，ERAP の最大の特徴は日英 2 言語のデータを収集していることである。論文に，著者自身による日本語アブストラクトと英語アブストラクトが含まれている場合は，両方をコーパスに加えた。また，英語のみのアブストラクトについては専門業者に日本語の対訳を作らせ，コーパスに加えた。同様に，日本語のみのアブストラクトについても英語の対訳をコーパスに収録している。

アブストラクトのベースになっているジャーナルの種別と収集したアブストラクトの本数は以下のとおりである。

表 1 ERAP 収集アブストラクトの本数とソース(予定含む)

	物理	化学	情報	計
海外誌	100	100	100	300
国内誌	100	100	100	300
国内博士論文	10	10	10	30
海外博士論文	20	20	20	60

地域(海外と国内), レベル(専門誌と博士論文), 分野(物理, 化学, 情報)の 3 観点を区別したうえで工学論文アブストラクトを日本語・英語の両方を切り口として分析できるのが ERAP の独自性である。

ERAP 所収のデータは, 英語については Stanford POS Tagger, 日本語は Janome で形態素解析を行っている。また, 筆者が談話単位(move)のコーディングを手作業で行っている(下図参照)。

図 1 ERAP のデータ構造

IPSJ-JNL601	セキュリティ対策導入にかかる時間とサイバーリスクレベル変動の関係から探る, 過剰なセキュリティ対策	Issues and Remedies for Excessive Security Controls Explored by Relationships between the Time taken to	0・・・タイト
IPSJ-JNL601	セキュリティ対策の実施には時間がかかるものがあるが、それによりサイバーリスクのレ	Cyber risk level sometimes rises for a limited time due to delay of the implementation of	1・・・背景
IPSJ-JNL601	一時的に大きくなるサイバーリスクのレベルを通常のサイバーリスク選好と比較して対応	If such a cyber risk level is compared with a normal cyber risk appetite, there are some	1・・・背景
IPSJ-JNL601	そこで、本論文では、セキュリティ対策の実施に時間がかかるとサイバーリスクのレベル	This paper proposes the issues of the implementation of excessive security controls	3・・・目的
IPSJ-JNL601	まず、サイバーリスクを作り出している要素として、サイバー空間に依存する企業価値、	First, the elements that create cyber risk: cyber accessible corporate value, attackers in	4・・・方法
IPSJ-JNL601	そして、定性分析モデルであるシステム・シンキングの理論を適用したうえ、それらが相	And then, how they interact each other and change cyber risk level is visualized by	4・・・方法
IPSJ-JNL601	また、実施に時間がかかるセキュリティ対策がサイバーリスクのレベルを一時的に大きく	The simulation on how cyber risk level rises for a limited time due to delay of the	方法／結果

2. ERAP Online 開発の現状と課題

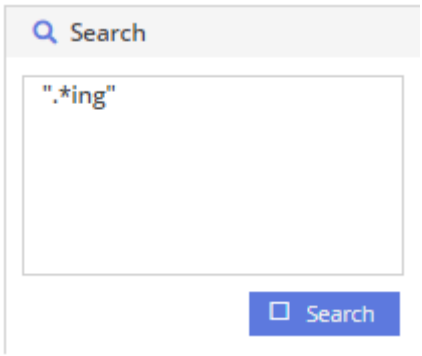
2.1 V0.1 の概要

すでに述べたように, 筆者の開発の狙いは日本人工学学生による自律的な学びの支援であり, この目的に照らしてみた場合, コーパスを作るだけでは十分でない。

そこで, コーパス言語学の知識を有さない学生であっても必要な検索が簡単に行えるよう, ERAP の専用検索システム(ERAP Online)の設計と開発を行っている(開発担当:Lago 言語研究所)。現在は V0.1 が試験稼働中である。

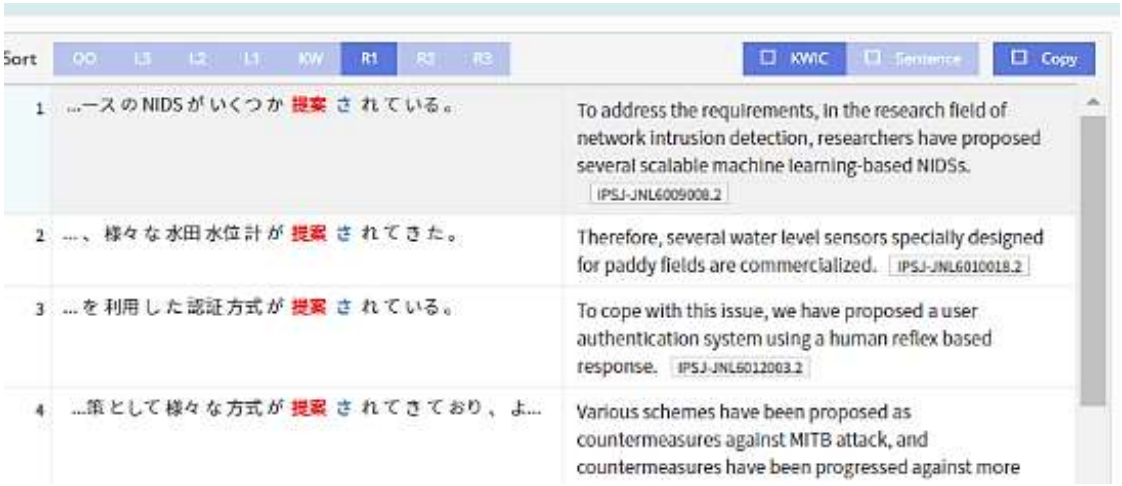
V0.1 では, 検索ボックスを用意しており, ユーザーは, 日本語または英語を入力する。正規表現にも対応しており, たとえば, 英語であれば, [. * ing] と指定することで, ing 形が一度に検索できる。また, 日本語であれば, [. * いる]と指定することで,「用いる」や「している」などが一度に検索できる。

図 2 検索ボックス(ing 形を指定した例)



検索語を入れると，当該語を含む用例が日英対称の形で表示される。下記は「提案」という日本語から検索を行った場合の結果である。

図 3 検索結果画面



左右位置をキーとしたソートのほか，全文表示(Sentence)にも対応している。また，出力結果をコピーすることもできる。

2.2 V0.1 の課題と対応

すでに述べたように，ERAP はユニークなデータであり，研究・教育両面での活用の可能性は大きいと思われるが，V0.1 を試行的に使用した工学学生にヒアリングしたところ，(1)正規表現のルールがむつかしい，(2)単語1語ではなく複数語で検索したい，といった要望が寄せられた。

そこで，V0.2 の開発にあたっては，正規表現についてはプルダウンからの選択で平易に使えるように検索系を改良する予定である。

一方，要望のあった複数語検索については，一般のコーパス言語学における必要性はわかるものの，ERAP のように用途が限定されたデータ検索において，1 語検索に比べて本当に有用な検索結果が得られるのかどうかははっきりしない。そこで，V0.2 の開発に先立ち，1 語検索と複数語

検索の場合の検索結果を比較し，その実際的有用性について検証を行うこととした。以下，検証内容と結果について報告する。

2. リサーチデザイン

2.1 研究課題

様々な分野の工学系ジャーナル論文要旨の言語特徴を分析した石川（2021）は，特に，情報系分野と化学系分野の英語アブストラクトにおいて，他の工学系分野とは異なる語の振る舞いが見られる可能性があることを指摘している。ここでは，ERAP Corpus から，情報系分野と化学系分野の日本語アブストラクトを取り上げ，それぞれ分野に特有の語彙・表現を抽出して，それらをどのように ERAP Online に組み込むとより効果的に自律学習を促すオンライン教材となるかを考える。現在のシステムは，単語の検索となっているので，今回は，語彙と連語に焦点を当てる。

RQ1. 単語単位の検索でわかることは何か。

RQ2. 連語単位（3-Gram）の検索でわかることは何か。

RQ3. これらを踏まえて現在のオンライン検索システムをどのように改善していくことが望ましいか。

2.2 コーパス・データ

ERAP Corpus の中から，日英対応の化学系論文アブストラクトと情報系論文アブストラクトをそれぞれ 100 本ずつ取り出し，本研究のデータとする。

表 2 データの概要

	情報系英語	情報系日本語	化学系英語	化学系日本語
語数	18,352	25,705	16,823	20,922
語種数	3,091	2,841	2,858	2,562

アブストラクトは，論文本体を読まなくても内容が分かるように書かれており，独立したテクストとなる。また，当該分野の専門家が流し読みをしても情報を正確に把握することを可能にするために，決められた字数・語数の中で，必要な情報が順序良く盛り込まれている。データ量は少ないが，専門分野での言語使用ルールに従った，「型」通りのテクストであることが期待できる。

なお，学生は，日本語を用いて検索を行うことが予想されるため，本研究では，日本語部分のみの分析とする。日本語の形態素解析には Mecab を，分析には CasualConc2.1.2 を用いる。また，比較には対数尤度比検定を用いている。

3. 結果と考察

ここでは，紙幅の関係上，談話単位についての分析は行わず，情報系・化学系のそれぞれの分野で特徴的に使用されている日本語の語彙と連語（3-Gram）のうち，上位 10 位を取り上げて比較検討を行うこととする。

3.1 RQ1 単語単位の検索でわかることは何か

まず，日本語データを用いて単語単位での分析を行い，分野別の特徴語を見てみよう。

表 3 情報・化学分野の日本語アブストラクトの特徴語

情報		化学	
特徴語	頻度（LL 値）	特徴語	頻度（LL 値）
者	135 (149.4)	膜	87 (140.5)
提案	194 (142.1)	粒子	84 (135.7)
情報	127 (140.1)	ガス	68 (109.8)
手法	153 (121.9)	液	68 (109.8)
する	599 (112.2)	濃度	66 (106.6)
攻撃	81 (95.8)	た	716 (103.4)
ユーザー	64 (75.3)	熱	68 (100.6)
学習	62 (73.3)	添加	57 (92.1)
通信	53 (62.7)	温度	57 (92.1)
な	298 (55.2)	層	62 (91.4)

情報と化学の各分野でそれぞれの特徴語がうまく抽出されていることが分かる。自律的語彙学習支援には有用なリストとなる可能性がありそうだ。しかし，ここからは，要旨の「型」や筆者が言いたい内容を十分にくみ取ることはできず，ライティングの学習へとつなげることは難しいように思える。また，形態素解析の知識のない工学系の大学院生にとって，「利用者」や「管理者」などの「者」を 1 語として検索しなければならないために，実際の検索が困難になっていることも推測できる。

3.1 RQ2. 連語単位（3-Gram）の検索でわかることは何か

次に，日本語データを分析して，分野別の特徴連語を見てみよう。

表 4 情報・化学分野の日本語アブストラクトの特徴連語（3 Gram）

情報		化学	
特徴連語	頻度（LL 値）	特徴連語	頻度（LL 値）
を提案する	68 (81.0)	検討した	41 (51.9)
手法を提案	30 (35.7)	ことがわかっ	25 (40.1)
本稿では	23 (27.4)	がわかった	24 (38.5)
論文では	42 (26.2)	について検討し	18 (28.9)
本論文で	41 (25.2)	を検討し	14 (22.4)
を実現する	21 (25.0)	得られた	24 (20.3)
することで	37 (21.4)	検討を行っ	11 (17.6)
の有効性	16 (19.1)	添加した	11 (17.6)
ている しかし	15 (17.9)	た。その結果	30 (16.7)
利用者が	15 (17.9)	を添加し	10 (16.0)

上記を見るだけで，情報分野においては，「本稿（論文）では，X（の手法）を提案することで， Y の有効性を示し，Z を実現する」ことが要旨の骨子であるのに対し，化学分野では，「X を添加して Y（について）検討した。その結果，Z がわかった（Z の結果が得られた）」と述べることが要旨の骨子となっていることが分かる。両分野においては，それぞれの研究の作法を反映し，要旨の執筆様式そのものにも質的な差が存在することが示唆される。

3.3 RQ3. 現在のオンライン検索システムをどのように改善していくことが望ましいか。

上記の結果より，ERAP Online システムの改良に向けては次の 2 点の追加が考えられる。

(1) 単語検索だけでは，要旨の骨子や主たる内容をくみ取ることができないために，連語検索または連語の提示を可能にする必要がある。

(2) 同じ工学系であっても，分野によって必要な表現は異なっており，分野別の検索・結果の提示ができるようにフィルターを活用する必要がある。

また，「本論文では」「本研究では」のように意味が近い表現も数多く見られる。それぞれに対応する英語をすべて提示するのではなく，頻出表現をまとめて提示することで学習負荷が軽減できる。そのため，学習者が単語や表現を自由に検索する方式だけではなく，N-Gram で抽出された頻出連語をまとめて表示し，そこから，「言いたいこと」に近い表現を選択するという方式を追加することも検討するべきであると考えている。

4. まとめと課題

本研究では，情報分野と化学分野における日本語アブストラクトを分析し，各分野で特徴的に使用されている語彙と連語を抽出した。結果から，単語単位の検索では分からなかった要旨の骨子や主たる内容を示す表現が提示できた。これらは日本人大学院生が「言いたいこと」である可能性が高く，自律的学習にも有用であると考えられる。ERAP Online に新たに連語検索や分野選択を組み込むことで，効果的な自立学習支援教材となる可能性がある。

謝辞

本研究は JSPS 科研費 19H01281 の助成を受けたものです。

引用文献

石川有香（2017）「English Vocabulary for Engineers 9000 の開発」『統計数理研究所共同研究リポート』*373-374*, 129-148.

石川有香（2020）「工学系大学院生を対象とした英語ニーズ調査―使用状況と自己評価の分析―」『言語文化学会論集』*53*, 201-215.

石川有香（2021）『ジャンルとしての工学英語』大学教育出版

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

N-grams at the Beginning of the Moves in the Results Section of Experimental Medical Research Articles

ISHII　Tatsuya (Kobe City College of Technology)
kcct-t-ishii@g.kobe-kosen.ac.jp
KAWAMOTO　Takeshi (Hiroshima University)
tkawamo@hiroshima-u.ac.jp

Abstract

Using a corpus based on move analysis of experimental medical research articles (approximately 1.5 million words in total), Ishii and Kawamoto (2020) focused on the behavior of adverbs and successfully identified 26 lexical phrases for the three moves in the Results section: (RM1) Introducing experiments, (RM2) Announcing results, and (RM3) Commenting on results. However, although cycles of these three moves were identified, no clarity was developed as to how the moves start and are connected. In this study, to identify the n-grams at the beginning of the three moves, we extracted and examined the first sentences of each move. After Imao (2021) extracted the first sentences of these three moves, using a wildcard, we copied and pasted them into an Excel document to divide them into independent words. Imao (2021) then counted the frequencies of n-grams. The observation of the n-grams led to a description of highly frequent phrases for starting and connecting moves—for example, “to determine” in (RM1), “we observed” in (RM2), and “taken together, these results” in (RM3). The study provides new insights for investigating a corpus based on move analysis.

Keywords

move analysis, n-gram, move cycle, experimental medical research articles

1. Introduction

The flow of discourse in research articles (RAs) is accomplished by moves, which contain several steps and typical expressions. To establish a narrative in a RA, writers must understand the typical expressions specific to the function of moves. Until now, move analysis has qualitatively revealed the functions of moves and steps in a particular discipline, referred to as “discourse units” and “rhetorical strategies,” respectively (Gray et al., 2020: 139). In contrast, corpus studies have quantitatively examined the language patterns in a particular area. Using a corpus based on a move analysis of experimental medical RAs, Ishii and Kawamoto (2020) defined the three moves in a typical Results section and described 26 lexical phrases with adverbs. Figure 1 summarizes the functions of the moves, steps, and examples of lexical phrases identified by Ishii and

Kawamoto (2020).

Figure 1 *Summary of Results Section Advocated by Ishii and Kawamoto (2020)*

RM1: Introducing Experiments	RM2: Announcing Results	RM3: Commenting Results
Step (1) Describing aims and purposes	Step (1) Highlighting important results	Step (1) Generalizing/interpreting results
e.g., We *first* set out to	e.g., *Interestingly*, we found that	e.g., *Collectively,* these findings indicate that
Step (2) Making hypotheses	Step (2) Showing additional or adversative results	Step (2) Emphasizing relationships
e.g., We *therefore* hypothesized that	e.g., *However*, we found that	e.g., bind *directly* to
Step (3) Listing procedures or methodological techniques	Step (3) Describing quantitative data	
e.g., *First*, we examined the effect of	e.g., was *significantly* higher than	

Because several experiments are conducted in experimental medical RAs, there are cycles of the three moves. However, it remains unclear as to how these moves start and are connected. Although moves are cycled, there is a slightly different function between the first round of (RM1) and the second and beyond round of (RM1). The first round of (RM1) was examined based on the hypothesis of the previous studies, while the second round and beyond round of experiments were conducted in response to the results and interpretation of the first round of experiment. Therefore, to understand how to make move cycles, it is necessary to investigate n-grams at the beginning of the first round of (RM1) and the second and beyond of (RM1) as well as (RM2) and (RM3). Thus, this study sought to answer four research questions (RQs):

(RQ1) What n-grams are highly used at the beginning of the first round of (RM1)?
(RQ2) What n-grams are highly used at the beginning of the second and beyond of (RM1)?
(RQ3) What n-grams are highly used at the beginning of (RM2)?
(RQ4) What n-grams are highly used at the beginning of (RM3)?

2. Method

2.1 Corpus Data

To investigate experimental medical RAs with an Introduction, Methods, Results, and Discussion structure, we excluded four articles that contained the Results and Discussion section from the corpus data created by Ishii and Kawamoto (2020), which included 300 RAs selected from 30 journals and published in 2014 (in total, 1,526,552 tokens). Moreover, the files of (RM1) were divided into those of the first round of (RM1) and of the second and beyond of (RM1). The corpus data of the Results section are show in Table 1.

Table 1 *Corpus Data of the Results section*

Corpus name	Function of moves	Tokens	Types
The first round of (RM1)	Introducing experiments	29,256	6,195
The second and beyond of (RM1)	Introducing experiments	237,778	18,778
(RM2)	Announcing results	369,723	22,985
(RM3)	Commenting results	80,786	8,789

2.2 Procedures

To answer the four RQs of this study, the following procedures were conducted.

1) Create text files for the first round of (RM1), the second and beyond of (RM1), (RM2), and (RM3).
2) Search the wildcard * using Imao (2021).
3) Arrange the concordance lines based on L1-R1-R2.
4) Copy the concordance lines into an Excel document.
5) Use a text import wizard and set the separator to space.
6) Delete the table and figure numbers.
7) Copy the vertical columns that contain the necessary number of words into a text file.
8) Produce the list of n-grams using Imao (2021).
9) Categorize the n-grams based on parts of speech.

3. Results

3.1 Answers to (RQ1)

The first round of (RM1) was defined as extending from the beginning of the Results section until the end of the first paragraph or the first round of (RM2). Of 300 RAs, four did not include (RM1), and 17 did not start with (RM1) in the Results section. Thus, 279 RAs started with (RM1) and were analyzed to produce 1-grams to 5-grams at the beginning of the first round of (RM1). The following figure summarizes these findings. The to-infinitive and the pronoun “we” were important signals for starting the first round of (RM1).

Figure 2 *Summary of n-grams at the beginning of the first round of (RM1)*

Grammatical Category	Expressions	Tokens	Examples	Tokens	Grammatical Category	Expressions	Tokens	Examples	Tokens
Adverbial Phrases	To	82	To investigate	15	Subjects	We	62	We first	7
			To identify	11				We performed	4
			To determine	5				We examined	3
			To examine	5				We used	3
			To explore	5				We analyzed	2
			To address	3				We investigated	2
			To evaluate	3				We measured	2
			To gain	3				We previously	5
			To study	3				We have previously shown that	2
	In order to	4						We set out to	2
	Using	3						We sought to identify	2
	Given	1	Given that	1		Our	6		2
Adverbs	Previously	2	Previously, we	2		It	4	It is	
						Previous studies	2		

3.2 Answers to (RQ2)

To answer (RQ2), we analyzed the 4565 first sentences at the beginning of the second and beyond of (RM1). As shown in Figure 3, in addition to the first round of (RM1), to-infinitives and the pronoun "we" were important signals for starting the second and beyond of (RM1). Since the experiments of the second and beyond of (RM1) were conducted based on the result of the previous experiment, the phrase "Having established" or the conjunctions "Because" and "Since" were used to confirm the interpretation of the previous experiment.

Figure 3 *Summary of n-grams at the beginning of* the second and beyond of *(RM1)*

Grammatical Category	Expressions	Tokens	Examples	Tokens	Grammatical Category	Expressions	Tokens	Examples	Tokens
Adverbial Phrases	To	1135	To determine	185	Adverbial Clauses	Because	117		
			To test	105		Since	49		
			To investigate	73		Although	39		
			To confirm	67		If	19		
			To assess	58	Subjects	We	994	We next	291
			To examine	47				We also	106
			To identify	34				We then	106
			To evaluate	24				We used	37
			To validate	20				We further	33
			To understand	19				We therefore	26
			To explore	19				We first	21
			To address	18				We investigated	19
			To verify	17				We tested	19
			To gain	17				We performed	17
			To extend	16				We examined	14
			To directly	15				We hypothesized	14
			To better	12				We analyzed	13
			To further	152				We reasoned	12
	In order to	36						We focused	10
	using	17						We previously	6
	Given	63	Given that	31				We have previously	5
	Having	38	Having established	15		Our	57	Our data	12
			Having shown	5				Our results	7
	Based on	18						Our observations	4
	On the basis of	11						Our finding	3
	In addition	33	In addition to	18				Our previous	3
			In addition,	14		It	34	It has	13
Adverbs	Next	128	Next(,) we	113				It is	11
	Finally	69	Finally(,) we	55				It was	7
	Furthermore	21	Furthermore(,) we	12					
	Therefore	17	Therefore(,) we	14					
	Thus	13	Thus(,) we	7					
	Previously	12	Previously(,) we	7					
	Moreover	11	Moreover(,) we	8					

3.3 Answers to RQ3

To answer (RQ3), we analyzed the 6026 first sentences at the beginning of (RM2). As shown in Figure 4, the pronoun "we" was followed by past-tense objective verbs such as "found" and "observed." Moreover, adverbs modifying the sentences can be seen as the signal of (RM2).

Figure 4 *Summary of n-grams at the beginning of (RM2)*

Grammatical Category	Expressions	Tokens	Examples	Tokens	Grammatical Category	Expressions	Tokens	Examples	Tokens
Adverbial Phrases	As	235	As expected	86	Adverbs	Indeed	93	Indeed(,) we	16
			As shown in Figure _	58		Interestingly	95	Interestingly(,) we	17
	Consistent with	137				Notably	55	Notably(,) we	10
	In contrast	66	In contrast,	31		However	84	However(,) we	16
			In contrast to	25		Strikingly	29	Strikingly(,) we	3
	In agreement with	15				Importantly	28	Importantly(,) we	3
	In line with	10				Remarkably	25	Remarkably we	9
	Using	64				Moreover	23	Moreover(,) we	2
	After	51				Surprisingly	21	Surprisingly(,) we	6
	In addition	40	In addition,	23		Similarly	18	Similarly(,) we	2
			In addition to	17		Consistently	12	Consistently(,) we	5
	Compared	35	Compared with	21		Conversely	10		
			Compared to	14		There	32	There was no	9
Adverbial Clauses	Although	93						There was a	8
	When	72	When we	16				There was an	4
	While	32			Conjunctions	and	237	and found	100
	Whereas	20						and we	14
Subjects	We	585	We found	282				and confirmed	13
			We observed	81					
			We identified	42					
	This	114	This analysis revealed	22					
			This analysis showed	4					
			This revealed	8					
			This resulted in	7					
	Analysis	34	Analysis of	33					
	Figure	29	Figure _ shows	19					

3.4 Answers to (RQ4)

To answer (RQ4), we analyzed the 4565 first sentences at the beginning of (RM3). As shown in Figure 5, when the result was interpreted, -ing forms or non-defining relative clauses with the verbs "suggest" and "indicate" were used. In contrast, when the interpretation was based on several results, the adverbial phrases "taken together" and "in summary" or the adverbs "together" and "collectively" were used.

Figure 5 *Summary of n-grams at the beginning of (RM3)*

Grammatical Category	Expressions	Tokens	Examples	Tokens	Grammatical Category	Expressions	Tokens	Examples	Tokens
Adverbial Phrases	Taken together	156	Taken together(,) these results	49	Subjects	These results	265	These results indicate	60
			Taken together(,) these data	40				These results suggest	57
	In summary	25	In summary(,) our	6				These results demonstrate	21
			In summary(,) we	5		These data	196	These data suggest	65
Adverbs	Together	160	Together(,) these data	60				These data indicate	36
			Together(,) these results	44		These findings	68	These findings suggest	22
	Collectively	78	Collectively(,) these results	24				These findings indicate	11
			Collectively(,) these data	21		These observations	24	These observations suggest	9
	Thus	289	Thus(,) we conclude that	6				These observations indicate	3
	Therefore	88	Therefore(,) we	7		This	243	This suggests	59
			Therefore(,) our	5				This indicates	15
			Therefore(,) it is	5				This is consistent with	14
-ing participle clauses	, suggesting	442	, suggesting that	326				This result	33
	, indicating	299	, indicating that	230				This observation	19
Non-defining relative clauses	, which	112	, which suggested	27				This finding	14
			, which indicated	27		We	56	We conclude that	3
			, which suggests	14					
			, which is consistent with	8					

4. Summary

Using Imao (2021), we have shown the n-grams at the beginning of the first round of (RM1), the second and beyond of (RM1), (RM2) and (RM3). These n-grams are used to create a story for the Results section of experimental medical RAs, referred to as moves cycles. The following examples show a template for the cycles of the three moves.

(RM1) To investigate

(RM2) We found that

(RM3) Taken together, these data show that

(RM1) We next

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 21K13011.

Bibliography

Gray, B., Cotos, E., & Smith, J. (2020). Combining rhetorical move analysis with multi-dimensional analysis: Research writing across disciplines. In U Römer, V. Cortes, & E. Friginal (Eds.), *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise*. (pp.137-168). John Benjamins Publishing Company.

Ishii, T. & Kawamoto, T. (2020). The behavior of adverbs in the results sections of experimental medical research articles: A corpus-based move analysis. *English Corpus Studies*, *27*, 23-52.

Imao, Y. (2021). CasualConc (Version 2.1.6) [Computer Software].
URL: https://sites.google.com/site/casualconc/download

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

生化学英語学術論文のための学術語彙リスト

清水　眞（東京理科大学）

makoto@rs.tus.ac.jp

村田　真樹（鳥取大学）

Academic Word Lists for Biological Chemistry

SHIMIZU, Makoto (Tokyo University of Science)
MURATA, Masaki (Tottori University)

Abstract

Although studying English presents considerable difficulties for Japanese university students, in this time of globalism, students are expected to read and write academic articles in English.　In order to solve this problem, we have made academic lists of organic chemistry and physical chemistry. In this study, we compile a corpus of a biological chemistry journal, made an academic list of biological chemistry, and discuss the characteristics of biological chemistry terms.

Keywords
学術語彙リスト、基本語、特殊学術目的の英語、一般学術目的の英語

1. はじめに

日本の大学における教養の英語教育の問題点は、従来から色々と指摘されて来た。日本人全体の英語の習熟度の低さ（石川　pers com によれば、日本人の 85%が CEFER の A1）、大学入学時における英語の習熟度の低さ（日本人大学生の TOEIC の平均点は 445 点）、必修の英語の履修は 1、2 年次の4コマに過ぎない、などである。にもかかわらず、研究室に配属された学部の 4 年生は、英語で書かれた学術論文を読み、理解することが求められている。この問題に対する解決策はいくつかあるだろうが、本研究では、語彙に焦点をあてる。

学術目的の英語(English for Academic Purposes　以下 EAP)においては、Coxhead(2000)の Academic Word List（以下 AWL)が編纂されて以来、多くの学術語彙リストが作成された。一般学術目的の英語(English for General Academic Purposes　以下　EGAP)を提唱する

Coxhead に対して、Hyland & Tse (2007)は、分野別である特殊学術目的の英語(English for Specific Academic Purposes 以下 ESAP)を提唱した。Shimizu *et al.* (2018)は、Hyland & Tse にならい、有機化学論文誌、物理化学論文誌に掲載された論文からコーパスを編纂し、有機化学論文のための学術語彙リスト(JACS)、物理化学論文のための学術語彙リスト(JPC)を作成した。この研究では、2016 年に発行された生化学論文誌である Journal of Biological Chemistry(以下 JBC)からコーパスを編纂し、生化学論文のための学術語彙リストを作成する。このリストと有機化学、物理化学のリストとの比較を行う。

2. 方法

学術語彙リストを作成するにあたり、データベースとしてコーパスを作成した。化学分野のなかの、生化学という下位分野に特化した。2016 年に発行された生化学論文誌から論文をそれぞれ 100 本選び、コーパスを作成した。

これらのコーパスから、すべての語のトークンをレンマ化し、品詞別にカウントしたリストを自動的に生成した。用いたソフトウェアは、Schmid (1997)のtagger、Charniak(2000)のCharniak Reranking Parserである。このリストから、以下を人手で削除した。

1) 固有名詞 Shape、University など
2) アルファベット 1 文字 b、c など（a は例外）
3) アルファベット 1 文字+. p.、S.、 E.など
4) アルファベット 1 文字+/ g/など、
5) 所有格を示す's
6) アラビア数字 20, 0.05 など
7) 単位記号 kg、mm、Hz
8) 略語 log、pH、DNA、HIV など
9) 元素記号 Fe、Na など
10) 記号 (、}、:、=、%、/など
11) html タグ &#8211、0x830xcam など

残ったものから、論文100本中生起度が20未満のものを除き、生化学のための学術語彙リスト、JBC2016を作成した。基本語として、JACET基礎2000語（大学英語教育学会基本語改定特別委員会(2015:30)参照。実際には2188語。以下基本語）のワードファミリーを用いた。生化学の専門家（JBCをメインの学術誌として使用する研究室の教員）に、化学、生化学の訳語があるものをチェックしてもらった。

3. 結果

JBC2016 は、タイプ数が 1880、内訳は、名詞 919、動詞 320、形容詞 387、副詞 146、その他 108 であった。JBC2016 のコーパスの総トークンは推定約 38 万であり、1880 タイプの語が約 92%をカヴァーするという計算結果が出た。うち基本語は 776 語であるが、その中には化学などの専門用語であるものも約 50 語含まれている。

名詞上位 10 語は、cell「細胞」(6555 トークン)、protein「タンパク質」(2778)、expression「発現」(1674)、activity「活性」(1127)、mouse「マウス」(991)、substrate「基質」(968)、study「研究」(927)、antibody「抗体」(918)、effect「結果」(888)、complex「複合体」(844)である。このうち、cell 、antibody、は生物学の、protein、activity、substrate、は生化学、expression は遺伝学の専門用語である。mouse は辞書に専門用語と記載されてはいないが、生物、生化学等の実験でよく用いられるので、「準専門用語」と呼ぶことができるかもしれない。「ハツカネズミ」と訳さず、「マウス」と訳すのが根拠のひとつである。Study、effect は AWL に記載されている。expression 、activity は基本語にリストアップされているが、生化学の専門家は、生化学の専門用語として、訳語を記載した。このように、基本語で、生化学の専門家が化学、生化学などの用語と認定したものは、約 50 タイプある。

動詞上位 10 語は、contribute「寄与する」(2017 トークン)、indicate「示す」(629)、induce「〈タンパク質・酵素を〉誘導する」(586)、regulate「規定する」(2778)、express「発現する」(524)、reduce「還元する」(492)、incubate「〈細菌・細胞などを〉培養する」(445)、analyze「…を(元素などに)分解する」(444)、inhibit「…の化学反応を抑制する」(439)、demonstrate「…を(実例・標本・実験などによって)(具体的に)説明する, 例示[例証]する」(388)である。incubate、demonstrate は実験、analyze、inhibit は化学、induce は生化学、express、reduce は生物の専門用語として辞書に記載されている。demonstrate は AWL にリストアップされている。

形容詞上位 10 語は、human「人の」(634 トークン)、mutant「突然変異による」(495)、binding「結合の」(405)、specific「種に特有の」(298)、cellular「細胞の」(289)、endothelial「血管内皮の」(246)、active「活性な」(243)、consistent「無矛盾の」(243)、endogenous「内在性の」(205)、molecular「分子の」(179)である。active 、molecular は化学、specific、cellular、endothelial、endogenous は生物学、mutant は遺伝学、binding は物理学、consistent は統計学の用語として辞書に記載されている。human は専門用語とは言えないが、生化学の研究対象の一つが人体であるため、これも。準専門用語と呼ぶことができるかもしれない。

副詞の上位 10 語は、significantly「有意に」(199 トークン)、further「さらに」(164)、furthermore「その上に, さらに」(131)、in vitro「生体外の」(97)、specifically「特に, とりわけ」(100)、similarly「同様に」(67)、overnight「一晩中」(67)、briefly「手短に言えば」(64)、stably「安定して」(64)、prior to「…より前に, に先立って」(61)である。in vitro は生物学、stably は化学の専門用語として辞書に記載されている。significantly 、further 、specifically 、similarly 、prior to は AWL に記載されている。furthermore、briefly は AWL に記載されていない。overnight については、専門家から、「生化学では, 12 時間以上は厳密に決める必要は

なく，そのような時は overnight と表記する。頻繁に使われる単語である。」というコメントをもらった。準専門用語と呼ぶことができるかもしれない。

その他で生起度が高いものに、whereas(368 トークン)、such as(330)、in response to(221)、due to(220)、according to(210)、as well as(205)、because of(171)、versus(143)がある。

これに対して、有機化学学術論文においては、名詞上位 10 語は、figure「図」(1064トークン)、datum「データ」(742)、protein「タンパク質」(722)、molecule「分子」(495)、complex「錯体」(480)、substrate「基質」(459)、pH(438)、NMR(426)、spectrum「スペクトル」(425)、mechanism「メカニズム」(419)である。

動詞上位 10 語は、observe「観察する」(595 トークン)、support「支持する」(519)、determine「結論を下す」(425)、contain「含む」(419)、bind「結合させる」(415)、suggest「示唆する」(405)、form「形成する」(378)、obtain「～を得る」(377)、provide「供給する」(364)、indicate「示す」(335)である。

形容詞上位 10 語は、experimental「実験の」(480 トークン)、binding「結合の」(362)、molecular「分子の」(346)、structural「構造の」(298)、catalytic「触媒の」(269)、solvent「溶剤の」(251)、significant「有意の」(223)、consistent「無矛盾の」(211)、available「有効な」(205)、radical「基の」(179)である。

副詞上位 10 語は、respectively「それぞれ」(199 トークン)、previously「前に」(164)、significantly「有意に」(346)、appropriately「適切に」(298)、relatively「比較的に」(97)、typically「典型的には」(251)、furthermore「さらに」(67)、readily「容易に」(64)、experimentally「実験的には」(64)、negatively「否定的に」(61)である。

その他は、whereas(101 トークン)、such as(198)、in response to(20)、due to(280)、according to(88)、as well as(144)、because of(78)、versus(75)である。

物理化学学術論文においては、名詞上位 10 語は、figure「図」(1697 トークン)、molecule「分子」(1319)、atom「原子」(1062)、particle「粒子」(892)、calculation「計算」(890)、datum「データ」(889)、spectrum「スペクトル」(810)、interaction「相互作用」(778)、simulation「シミュレーション」(770)、parameter「パラメータ」(754)である。

動詞上位 10 語は、indicate「示す」(658 トークン)、occur「起こる」(506)、define「定義する」(284)、bind「結合させる」(270)、coordinate「配位結合させる」(249)、assume「推定する」(234)、yield「生じる」(198)、compute「計算する」(184)、associate「…を(…と)会合させる」(364)、assign「〈属性・役割・名称・構造などを〉(…が)持っているとする」(175)である。

形容詞上位 10 語は、experimental「実験の」(894 トークン)、molecular「分子の」(596)、constant「一定な」(476)、solvent「溶媒の」(446)、initial「初期の」(409)、consistent「無矛盾の」(367)、significant「有意な」(333)、corresponding「対応する」(332)、electronic「電子の」(299)、vibrational「振動の」(264)である。

副詞上位 10 語は、respectively「それぞれ」(444 トークン)、thus「」(415)、significantly「有意に」(217)、appropriately「およそ」(188)、slightly「わずかに」(187)、experimentally「実験的

に」(182)、typically「典型的に」(142)、strongly「強く」(139)、hence「したがって」(110)、essentially「本質的には」(86)である。

その他は、whereas(164 トークン)、such as(255)、in response to(11)、due to(393)、according to(118)、as well as(170)、because of(156)、versus(121)である。

4.考察

基本語のうち、zoo、diary などの語が、生化学の学術論文に生起しないことは予想がつくが、人称代名詞でも生起しないものがあることを指摘しておきたい。1 人称単数、2 人称、3 人称単数男性形、女性形は全く生起しない。また、再帰代名詞は、itself を除いて生起しない。基本語のうち学術論文に生起するものは、約 800 語である。

生化学、有機化学、物理化学の学術語彙リストを比較して、まず気がつくことは、有機化学、物理化学の学術語彙リストが似ているのに対して、生化学の学術語彙リストはそれらとかなり異なるということである。例えば、有機化学、物理化学の学術語彙リストの名詞上位 10 語に、figure、datum、molecule、spectrum が共通して入っているのに対し、生化学の学術語彙リストには、共通するものがない。また、形容詞上位 10 語で、experimental、molecular、solvent、significant、consistent が共通して入っているのに対し、生化学の学術語彙リストで共通するものは、molecular と consistent のみである。副詞上位 10 語では、respectively、significantly、typically、experimentally が共通して入っているのに対し、生化学と有機化学の学術語彙リストで共通するものは、significantly と furthermore のみである。

上位 10 語のみしか示すことができなかったが、この傾向はそれぞれの学術語彙リスト全体に見られる傾向である。同じ化学という分野でも、生化学、有機化学、物理化学では、語のタイプ、トークンが異なる。また、有機化学と物理化学の方が、有機化学と生化学、物理化学と生化学よりも、生起度において類似しているというように思われる。

5. まとめ

以上、学部の 4 年生が、英語で書かれた学術論文を読み、理解することを目的とした学術語彙リストを作成のために、生化学論文誌のコーパスを編纂し、生化学論文のための学術語彙リストを作成した。また、有機化学論文のための学術語彙リスト、物理化学論文のための学術語彙リストと比較、生化学論文の専門用語と特性を論じた。今後、学術語彙リストが実際にどのくらい有用であるか検証するため、インタビュー、実験を行いたい。また、無機化学論文誌のコーパスを編纂、無機学論文のための学術語彙リストの作成を行いたい。

謝辞

化学、生化学についてご指導くださった、東京理科大学理学研究科下仲基之先生に感謝したい。

引用文献

Charniak, Eugene. (2000). A maximum-entropy-inspired parser. *Proceedings of NAACL*, 132-139.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly* 34.1:2, 213-238.

大学英語教育学会基本語改定特別委員(編著)(2015)『大学英語教育学会基本語:新JACET8000』、桐原書店

Hyland, K. & Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly* 4.1.1:2, 235-253.

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. Jones & H. Somers (eds.) *New methods in language processing*. London: UCL Press.

Shimizu, M., Murata, M, & Ramonda, K. (2018), 'Teaching English for Chemistry at a Japanese University', *The Online Journal of Science and Technology* - July 2018Volume 8, Issue 3, http://www.tojsat.net/?pid=showissue&issueid=202

ウェブサイト

・Makoto Shimizu (Academic Word Lists)
http://www.rs.kagu.tus.ac.jp/makoto

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

高校英語指導における句動詞の扱い
―教科書とセンター試験の分析から―

堀家　利沙（神戸大学　大学院生）
209c109c@gsuite.kobe-u.ac.jp

Phrasal Verbs in English Teaching at Senior High Schools
A Corpus-based Analysis of Textbooks and National Center Test

HORIKE　Risa (Kobe University, Graduate Student)

Abstract

This study investigates phrasal verbs (PVs) presented in English materials for Japanese senior high school students. Six series of high school English textbooks and the last 10 years of the National Center Test were analyzed. The analyses showed that (1) textbooks and tests do not present sufficient number and types of PVs, (2) among 100 key PVs, only 12-69 types are presented in materials, and (3) textbooks deviate from standard modern English corpora in terms of the use of PVs.

Keywords

句動詞，日本人高校生，検定教科書，大学入試センター試験

1. はじめに

句動詞は，英語学習者にとって特に習得が困難な学習項目の 1 つである。堀家（in press）では，既存の大学生英作文コーパスと自作の高校生英作文コーパスを用い，日本人高校生，日本人大学生，英語母語話者の句動詞使用を比較したが，その結果，学習者の句動詞使用量は英語母語話者より少なく，使用される句動詞の種類も制約されていることが明らかになった。

では，こうした差異は何に起因するのであろうか。原因の一つは，学習者にとっての英語インプット，つまりは，学習者が使用している教材の内容に問題が残されている可能性である。この点を検証するため，本研究では，検定教科書とセンター試験（英語）をコーパス化し，これらを 2 種の現代英語コーパス（British National Corpus：BNC，Corpus of Contemporary American English：COCA）と比較することで，両者の乖離を明らかにし，教材に不足している句動詞の実態の解明を目指す。なお，センター試験は，本来，教材を意図したものではないが，石川（2019）の指摘にもあるように，高校までで学習すべき内容が集約されていることと，多くの高校でセンター試験問題を利用した指導が行われていることをふまえ，本研究では教科書と併せて分析対象とする。

2. 先行研究

ここでは，インプットが学習者の句動詞使用実態や理解度に与える影響について調査した 2 つ

の先行研究を紹介したい。句動詞理解度に焦点を当てた Negishi, Tono, & Fujita（2012）の研究では，日本人高校生および大学生を対象とし，句動詞 100 種の理解度試験を実施し，併せて中学校英語教科書コーパスの分析を行うことで，English Vocabulary Profile（以下，EVP）が句動詞に対して付与するヨーロッパ言語共通参照枠（以下，CEFR）レベルの妥当性の検証が行われている。同研究では，教科書において極めて出現頻度が低い句動詞（例：leave behind）や，教科書で扱われにくいトピックに関連する句動詞（例：split up）などは，EVP が付与する CEFR レベル以上に学習が困難である一方で，カタカナで借用語として定着している句動詞（例：knock out）や意味的透明性が高い句動詞は比較的学習の定着に繋がりやすいことが示唆されている。

石井（2018）は，日本人 EFL 学習者の話し言葉コーパスと中高等学校英語教科書コーパスを分析し，教科書の句動詞使用は，英語習熟度が中上級レベル（CEFR：B1.2, B2）の学習者の話し言葉における句動詞使用量とほぼ同程度だが，質的に見ると学習者と教科書の句動詞使用傾向には乖離が見られることを指摘している。

3. リサーチデザイン

3.1 研究目的と研究設問

前述のように，本研究は，教科書およびセンター試験コーパスと，現代英語コーパスにおける句動詞の出現状況を比較し，高校生のインプットにおける句動詞提示の課題を探ることを目指す。なお，ここで留意すべきは，検定教科書に，おおむね，初級用・中級用・上級用という 3 段階の区別が存在することである。また，学校の実情によって，センター試験を指導に利用している場合も，そうでない場合も存在する。つまり，高校生が接触する英語インプットについては，(a) 初級教科書のみ，(b) 中級教科書のみ，(c) 上級教科書のみ，(d) 上級教科書とセンター試験の併用，という 4 条件を区別して議論することが重要になる。これらをふまえ，以下の 3 つの研究設問を設けた。

RQ1：4 種のインプット条件で提示される句動詞の量（トークン・タイプ）は，現代英語コーパスと比較して，適切か。

RQ2：4 種のインプット条件において，重要句動詞 100 種はどの程度カバーされているか。

RQ3：句動詞使用の点で，教科書・センター試験・現代英語はどのような関係にあるか。また，それぞれを特徴づける句動詞は何か。

3.2 分析対象句動詞と使用するコーパス

本研究では，Gardner & Davies（2007）の枠組みをもとにし，動詞 20 種（go, come, take, get, set, carry, turn, bring, look, put, pick, make, point, sit, find, give, work, break, hold, move），不変化詞 8 種（out, up, on, back, down, in, over, off）の組み合わせにより構成される計 160 種の句動詞を分析対象としている。同研究において，これらは，BNC に出現する全句動詞の半数以上を占めることが報告されている。

現代英語としては，1 億語のイギリス英語コーパス，BNC と 10 億語の現代アメリカ英語コーパス，COCA の 2 種を参照する。各々のジャンル構成は異なるが，ここでは，両者に共通する書き言葉 4 ジャンル（フィクション，雑誌，新聞，学術）のみを比較対象とする。

教科書コーパスについては，2 つの出版社（東京書籍，三省堂）が刊行する初級，中級，上級の「コミュニケーション英語Ⅰ～Ⅲ」，計 6 シリーズ（初級 All aboard，Vista；中級 Power on, My way；上級 Prominence，Crown）を分析対象とした。各課の本文のみを入力範囲とし，独自に電子化して構築した 3 種のレベル別教科書コーパスの総語数は，初級が 10,842 語，中級が 28,184 語，上級が 53,620 語となっている。

センター試験コーパスについては，過去 10 年分（2011～2020 年）の本試験および追試験の中で，長文問題にあたる第 5，6 問を分析対象としている。本文のみを電子化し，構築したセンター試験コーパスの総語数は，51,117 語となっている。

3.3 手法

まず，RQ1 では 160 種の句動詞について，総出現頻度（トークン数）と，実際に出現している種別数（タイプ数）を調査する。トークン数については，各資料の総語数が異なることから，10,000 語当たりで補正する。タイプ数については，補正は行わない。

続く RQ2 では，160 種の中から筆者が別途実施した調査で特定した重要句動詞 100 種を分析対象とする。これは，BNC と COCA の各 5 ジャンル（話し言葉，フィクション，雑誌，新聞，学術）における出現頻度（全 10 種）を主成分得点化し，得点上位 100 種を選んだものである。なお，上位 5 種は，go on，set up，pick up，come back，go back である。

最後に，RQ3 では，第 1 アイテムを句動詞（全 100 種），第 2 アイテムをテキストタイプ（教科書 3 種・センター試験・BNC・COCA の全 6 種）とする頻度表を用意し，対応分析を実施し，得られた散布図を質的に解釈する。

4. 結果と考察

4.1 RQ1 句動詞の提示状況

語彙習得において，とくに偶発的学習（incidental learning）を可能にするには，圧倒的に大量の接触機会を持つ必要がある（松沢，2010）。日本人学習者の場合，教材以外の英語接触機会がほとんどないことをふまえると，教材では，BNC や COCA を少なくとも下回らない程度の句動詞が提示されていることが望ましい。この点をふまえ，160 種の句動詞の使用量について調査したところ，以下の結果を得た。

図 1 160 種の句動詞の総出現頻度（トークン数）

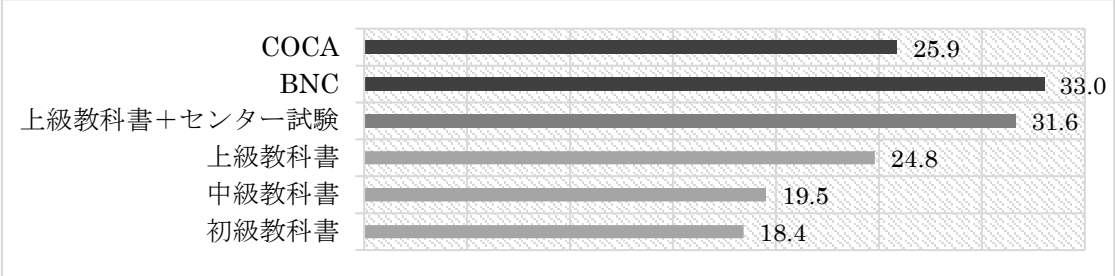
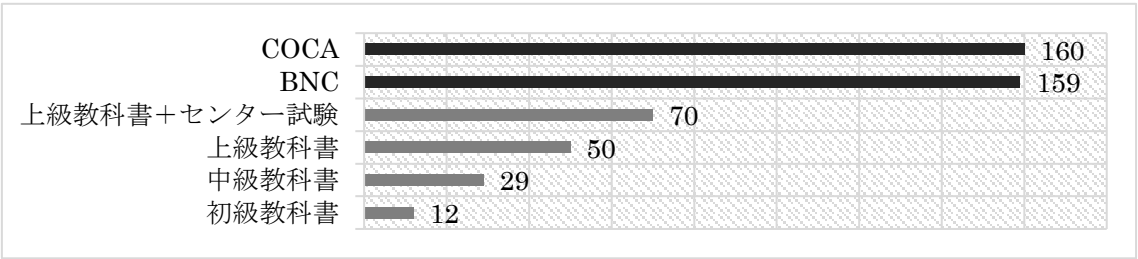


図 2 句動詞 160 種のうち出現している種別数（タイプ数）



まず，トークン数について，句動詞使用量がより多い BNC を基準とし，Fisher の正確確率検定を用いて，4 種のインプット条件における使用量と比較したところ，上級教科書とセンター試験を併せて使用した場合を除き，句動詞インプット量が少ないことが確認された（初級 p<.05, OR=0.56；中級 p<.05, OR=0.59；上級 p<.05, OR=0.75，上級＋センター試験 p=0.44, OR=0.96）。

次に，タイプ数については，BNC／COCA においては，調査対象の 160 種のほぼすべてが1回以上出現していたのに対し，教材に出現する句動詞のタイプ数は 12～70 と圧倒的に制限的であることが示された。使用種類数がより多い COCA を基準に実施した Fisher の正確確率検定でも，4 種のインプット条件すべての場合において少なく，有意差が認められた。

以上より，教材に提示される句動詞は，4 条件いずれの場合であっても，総頻度・種別数のいずれにおいても不十分であることが実証された。

4.2 RQ2 重要句動詞 100 種のカバー状況

前節では，相対的に低頻度のものも含む全 160 種の句動詞のうち，教材で提示される句動詞の種別数が 12～70 と圧倒的に少ないことが示された。では，現代英語の様々なジャンルで幅広く多用される重要 100 種に限ってみた場合，実態はどうであろうか。調査により，以下の結果を得た。

表 1 重要句動詞 100 種のカバー状況と出現なしの重要句動詞

インプット	出現	非出現	非出現の例（重要度ランク上位 5 種）
初級教科書	12 種	88 種	set up (2 位)，go back (5 位)，point out (6 位) come up (7 位)，find out (8 位)
中級教科書	28 種	72 種	come in（10 位），turn out（12 位），take on （15 位) come on（18 位），go down（19 位）
上級教科書	50 種	50 種	come on（18 位），sit down（21 位），get back（24 位） go in (26 位)，come down (27 位)
上級教科書 センター試験	69 種	31 種	come on（18 位），get back （24 位），go in (26 位) come down (27 位)，work in (29 位)

上表を概観すると，重要句動詞 100 種に限っても，教材で提示される句動詞の種別数は 12～69 にとどまることが確認された。このことは，多くの重要句動詞が，現行の高校生用教材において，まったく出現していないことを示唆する。もちろん，中学校教科書ですでに提示済みであるという判

断のもとにそれらが収録されていないという可能性はあるが，前述のように，語彙や連語の習得においてインプット量が不可欠であることをふまえれば，いわゆる中学レベルのものであっても高校教材で繰り返して提示されていることが望ましいと言える。もっとも，教材という観点から考えれば，学術論文でしか使用されないような高度にフォーマルなものが欠損していることは必ずしも問題とはいえない。一方，日常生活で多用されるインフォーマルな句動詞の欠損はより大きな問題になりうるだろう。そこで，欠損している句動詞のフォーマリティを検証することとしたい。

前述のように，筆者は，100 種の重要句動詞を選定する際に主成分分析を行ったが，第 1 主成分が各変数にプラスの負荷量を与える総合指標であったのに対し，第 2 主成分は学術ジャンルとフィクション・話し言葉ジャンルをわける軸，つまりはフォーマリティの高低の軸と解釈された。ここで，第 2 主成分のスコアをフォーマリティ指標（5 段階）とみなすと，表 1 に示した非出現かつ重要度の高い句動詞 20 例（重複を除くと 15 例）のフォーマリティレベルは以下のようになった。

表 2 インプットにおいて非出現の句動詞に見られるフォーマリティの特徴

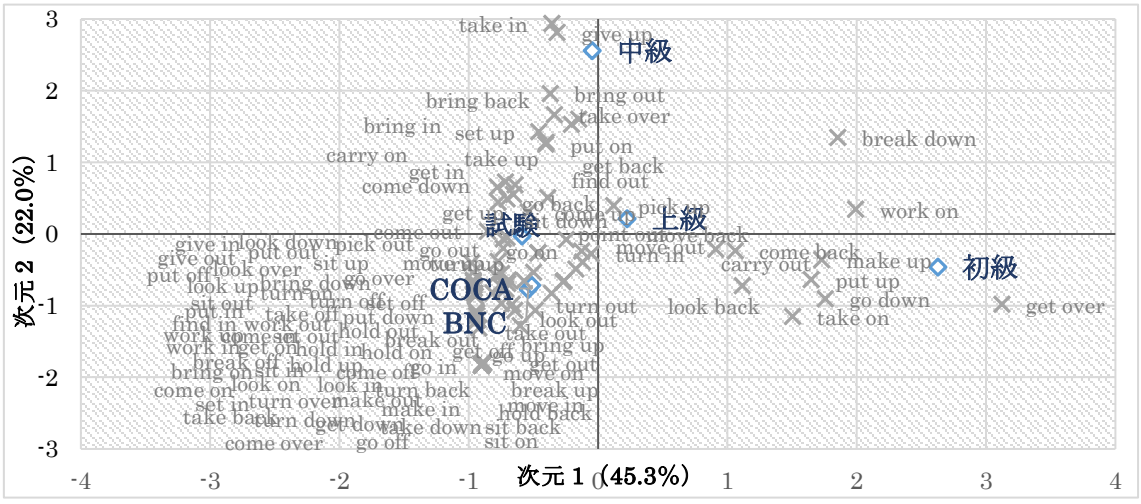
フォーマリティレベル（5 段階評価）	該当句動詞
高（レベル 5）	set up, point out, turn out, take on, work in
中（レベル 3～4）	該当なし
低（レベル 1～2）	find out, go down, get back, go in, come down, go back, come up, come in, come on, sit down

上記に示したもののうち，とくに，フォーマリティレベルの低いものについては，指導の現場で意図的に補っていくことが必要であろう。

4.3 RQ3 テキストタイプ間の関係性

対応分析を実施したところ，5 つの次元が取り出され，上位 2 次元で散布図を作成したところ，以下の結果が得られた。なお，第 2 次元までの累積寄与率は 67.3%であり，この散布図で元の分散の過半が集約されている。

図 3 対応分析に基づく散布図



まず，BNC と COCA の関係性に着目すると，散布図の中で両者がきわめて近い位置に付置されることが確認された。このことは，英・米問わず，現代英語において句動詞がある程度安定的なふるまいを見せることを示唆する。次に，6 種のテキストタイプの関係性に着目すると，第 1 軸上で教科書（初級・上級）とそれ以外が左右に分割されることから，教科書が現代英語とは，質的に異なる句動詞提示を行っている可能性が考えられる。一方で，広義の教材の中でもセンター試験は現代英語に近い位置にあることが確認された。

また，散布図を 4 つの象限に区別すると，BNC，COCA はいずれも第 3 象限に含まれている。つまりは，第 3 象限の句動詞は，現代英語を特徴づけるもので，かつ，教科書やセンター試験に十分に反映されていないものである。第 3 象限に付置された句動詞 71 種を前述のフォーマリティ指標をもとに区分すると，低 38%（come on, go on 等），中 51%（hold up, turn over 等），高 11%（turn out, point out 等）となり，下記に示すように，日常的場面で使用される句動詞も多い。

(1) "*Come on*. Don't give up so easily." （COCA_フィクション 2019）

(2) Plant collecting has been *going on* for thousands of years. (BNC_雑誌 1991)

以上のような句動詞は，日本人高校生にも優先的に指導すべきものであり，今後，教科書などにも積極的に採用することを検討する必要があるだろう。

5. まとめ

本研究では，高校生が量的，質的に充分な句動詞インプットを得られているかを検証すべく，調査を行った。高校英語インプットにおいて，句動詞トークン数，タイプ数は共に制限的であり，とくに日常的な場面で使用されるフォーマリティレベルの低い句動詞の扱いが手薄となっている可能性が示唆された。学習者の句動詞使用の幅を広げるためには，以上の句動詞インプットの特徴や制約をふまえたうえで，不足を補いながら，句動詞の多義性や用法を学習者が自ら発見できるような指導につなげていく必要があるだろう。今後，その一策としてデータ駆動型学習と重要句動詞リストを併用した教育実践も視野に入れ，研究を継続したい。

引用文献

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, *41*(2), 339-359.

堀家利沙(in press)「高大連携を志向した日本人英語学習者の基本動詞コロケーションの発達パタンのモデル化―学習者コーパスを使った研究―」『英検研究助成報告書』, *33*.

石井康毅(2018)「話し言葉コーパスと検定教科書に基づく日本人英語学習者の句動詞使用実態の分析」*Learner Corpus Studies in Asia and the World*(神戸大学), *3*, 101-119.

石川慎一郎(2019)「英語教育における連語：ターゲット・インプット・アウトプットの三元コーパス分析をふまえた English N-gram List for Japanese Learners of English(ENL-J)の開発と利用」. 仁科恭徳・吉村由佳・吉川祐介(編)『言語分析のフロンティア』(pp.32-47).金星堂.

松沢伸二(2010)「多量のインプットで英語力を育む」『Teaching English Now』, *19*.

Negishi, M., Tono, Y., & Fujita, Y. (2012). A validation study of the CEFR levels of phrasal verbs in the English vocabulary profile. *English Profile Journal*, *3*, 2-16.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

LDA Topic Modelling of Tennyson's Poetry

FUJITA Iku (University of Osaka, Graduate Student)
u256780k@ecs.osaka-u.ac.jp

Abstract

This study aims to provide an in-depth investigation of 66 epic and lyrical poems of the Victorian poet, Alfred Tennyson, which are over 1,000 words in length, using Latent Dirichlet Allocation topic modelling (henceforth LDA). Emerging LDA results have detected the latent topics hidden behind the prominent elements of the poems in the corpus, and the topics that appear in some works in common; of further interest are the latent connections between some works. In addition, this study discusses the possibility of detecting rhyming elements when LDA is run on poetry data as well as the issue of part-of-speech (POS) tagging on verse texts, suggested by the results of LDA in hindsight, and conceivable future approaches for addressing the issues.

Keywords

Alfred Tennyson, LDA, Topic Modelling, Poem

1. Introduction

Topic modelling is considered a promising approach in text mining (Meeks & Weingart, 2012). A number of studies have examined prose texts using topic modelling (Tabata, 2017; Kiyama, 2018; Huang, 2020). However, the application of topic modelling to poetry is still developing; few studies, apart from Rhody (2012), Navarro-Colorado (2018), Henrichs (2019), and Okabe (2019), have investigated the possibility of applying topic modelling to poetry. The use of the method still has room for improvement with regard to its application to a corpus of poems. This paper attempts to report emerging results of running Latent Dirichlet Allocation topic modelling (henceforth LDA) (Blei et al., 2003) and aims to provide an in-depth investigation using LDA on Alfred Tennyson's poetry.

2. Literature Review

One of the topic modelling algorithms, LDA (Blei et al., 2003), is now popularly used primarily for the examination of prose texts. In particular, Tabata (2017) and Kiyama (2018) both employ LDA on English texts, Dickens's novels and State of the Union Addresses, respectively, while Huang (2020) applies LDA to mystery fictions written in Chinese (to name but a few). On the other hand, few studies have investigated the

possibility of applying topic modelling to poetry. Rhody (2012) uses LDA to investigate whether the topic modelling method can detect the figurative language in 4,500 English poems. Navarro-Colorado (2018) and Henrichs (2019) apply LDA to sonnets, of which tokens lengths are approximately 100 to 200. Okabe (2019) uses the LDA on Emily Dickinson's poems. However, previous work in the literature tend to deal with short sonnets or lyrical poems, while Tennyson wrote a large number of epic poems that tend to have a larger number of word tokens in comparison with lyrical poems. Thus, the use of the topic modelling method still has room for improvement considering its application to a corpus of poems.

3. Data and Method

The data of this study is 66 epic and lyrical poems of the Victorian poet, Alfred Tennyson, which are over 1,000 words in length (Table 1). The corpus data were pre-processed as follows:

(1) Extracting only the body of the poems
(2) Separating the poems into 1,000–1,999-word files
(3) Renaming the file titles as the abbreviated poem titles and file numbers
(4) Removing function words and already-explicit items of which occurrences are limited to particular poems, including character names and honorific titles as stopwords

Table 1 The list of Tennyson's 66 poems

	Year of Publish	Titles of Poems	Token		Year of Publish	Titles of Poems	Token
1	1847	Princess	26,772	34	1829	Timbuctoo	1,903
2	1850	In Memoriam A H H	19,155	35	1842	St Simon Stylites	1,874
3	1859	Lancelot and Elaine-Idylls of the King	11,959	36	1833	OEnone	1,818
4	1872	THE ROUND TABLE Gareth and Lynette-Idylls of the King	10,851	37	1892	Akbars Dream	1,803
5	1855	Maud	10,280	38	1880	The Voyage of Maeldune	1,795
6	1864	Aylmers Field	9,461	39	1855	The Brook	1,790
7	1833	THE LOVERS TALE A FRAGMENT	8,738	40	1880	The Village Wife or The Entail	1,774
8	1857	Geraint and Enid-Idylls of the King	8,035	41	1889	Owd Roa	1,729
9	1857	Merlin and Vivien-Idylls of the King	8,024	42	1833	The Millers Daughter	1,728
10	1869	The Holy Grail-Idylls of the King	7,663	43	1885	The Spinsters Sweet-Arts	1,672
11	1862	Enoch Arden	7,531	44	1885	Tiresias	1,669
12	1857	The Marriage of Geraint-Idylls of the King	6,951	45	1880	Sir John Oldcastle Lord Cobham	1,635
13	1871	The Last Tournament-Idylls of the King	6,298	46	1881	Despair	1,563
14	1859	Guinevere-Idylls of the King	5,802	47	1880	The Northern Cobbler	1,555
15	1885	Balin and Balan-Idylls of the King	5,090	48	1842	Dora	1,504
16	1869	Pelleas and Ettarre-Idylls of the King	5,043	49	1842	Will Waterproofs Lyrical Monologue made at The Cock	1,488
17	1869	The Coming of Arthur-Idylls of the King	4,313	50	1827	THE OAK OF THE NORTH	1,480
18	1869	The Passing of Arthur-Idylls of the King	3,889	51	1842	The Vision of Sin	1,480
19	1889	The Ring	3,752	52	1833	The Lotos Eaters	1,433
20	1886	Locksley Hall Sixty Years After	3,355	53	1889	Romneys Remorse	1,383
21	1842	The Two Voices	3,199	54	1862	The Grandmother	1,370
22	1842	Morte dArthur	2,519	55	1830	Supposed Confessions OF A SECOND-RATE SENSITIVE MIND NOT IN UNITY WITH ITSELF	1,360
23	1880	The Sisters	2,357	56	1870	The Window or the Song of the Wrens	1,327
24	1842	Locksley Hall	2,326	57	1879	The Defence of Lucknow	1,311
25	1842	The Gardeners Daughter	2,282	58	1885	Tomorrow	1,298
26	1833	A Dream of Fair Women	2,272	59	1885	The Flight	1,289
27	1868	Lucretius	2,238	60	1889	Happy	1,282
28	1885	The Ancient Sage	2,173	61	1880	The First Quarrel	1,277
29	1842	OEnone	2,090	62	1878	The Revenge A Ballad of the Fleet	1,272
30	1832	The Palace of Art	2,069	63	1889	Demeter and Persephone	1,268
31	1880	Columbus	2,007	64	1851	Edwin Morris or The Lake	1,228
32	1885	The Wreck	1,979	65	1880	Rizpah	1,221
33	1852	Ode on the Death of the Duke of Wellington	1,937	66	1833	The Lady of Shalott	1,060

The Machine Learning for LanguagE Toolkit (MALLET) was installed to apply LDA to the Tennyson corpus. Before the number of topics was settled to 20, some experimental trials had been repeated with the number of topics changed from 10 to 200.

4. Results

The results of the LDA show that the topics hidden behind the prominent elements of the poems appeared in some works in common and the latent connections between some works were discovered.

Table 2 shows the hyperparameter alpha (α), the labels of topics, and the 20 keywords of each topic. The label of each topic is represented by the first three most strongly-contributing words to the given topic. A higher α value indicates a greater prevalence of the topic in the corpus. A very low α value shows the topic appears in a small number of texts, more often than not in a single text only.

Table 2 The list of topics with their keywords (20 keys)

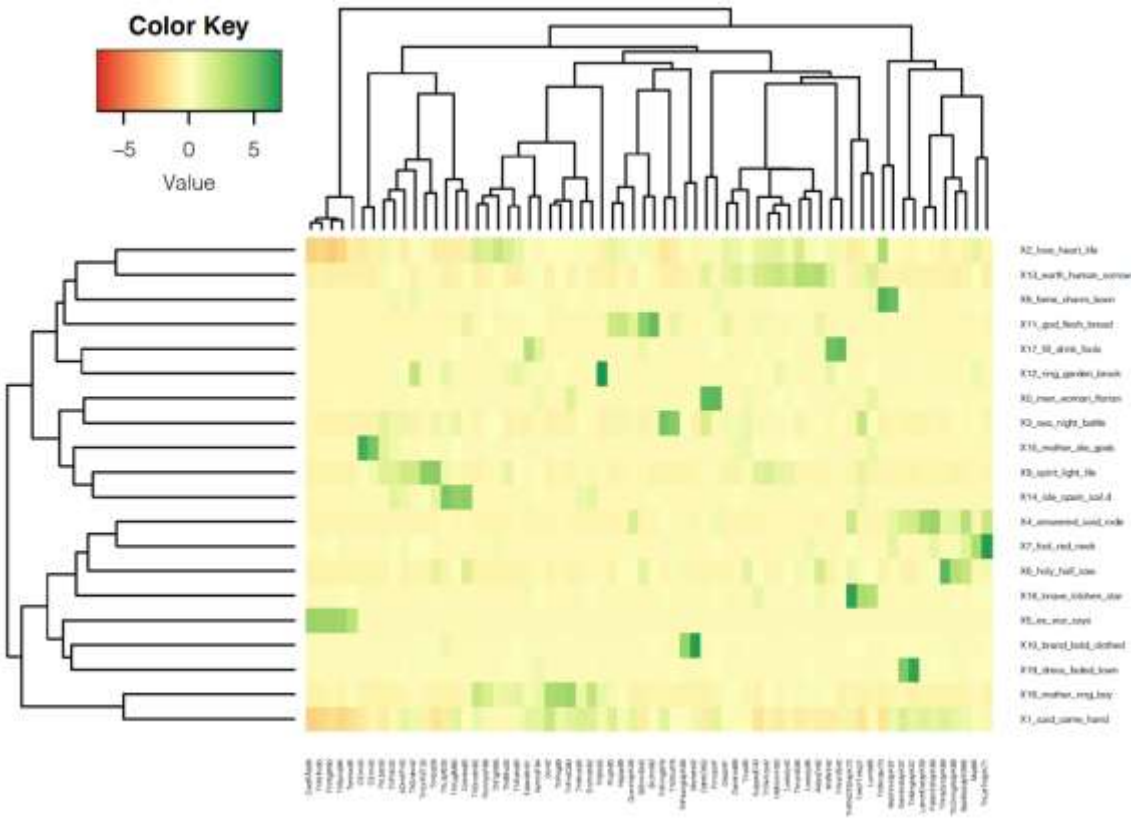
alpha	labels	keys (1–10)									
0.09065	0_men_woman_florian	men	woman	florian	iron	duty	honor	laws	college	spoke	common
1.79663	1_said_came_hand	said	came	hand	made	heard	face	left	rose	man	saw
2.28222	2_love_heart_life	love	heart	life	day	let	light	man	see	night	death
0.29520	3_sea_night_battle	sea	night	battle	war	land	hollow	curse	roof	dying	die
0.42260	4_answered_said_rode	answered	said	rode	horse	court	name	shield	maid	noble	spake
0.00564	5_es_wur_says	es	wur	says	niver	er	squire	owd	ma	ed	night
0.23551	6_holy_hall_saw	holy	hall	saw	city	heaven	seen	fire	earth	spake	son
0.10374	7_fool_red_neck	fool	red	neck	wood	hell	kind	oak	shore	broken	peace
0.14303	8_fame_charm_boon	fame	charm	boon	letter	answered	smiling	use	trust	south	fine
0.26351	9_spirit_light_life	spirit	light	life	place	hope	memory	stream	soul	seem'd	sound
0.11458	10_brand_bold_clothed	brand	bold	clothed	barge	spake	hilt	wind	mere	lake	sword
0.08154	11_god_flesh_bread	god	flesh	bread	worst	sin	saved	hope	heresy	priest	mercy
0.16291	12_ring_garden_brook	ring	garden	brook	seems	lake	rapt	whispers	fresh	hue	genial
0.24222	13_earth_human_sorrow	earth	human	sorrow	faith	race	grave	doubt	nature	lives	use
0.07544	14_isle_spain_sail'd	isle	spain	sail'd	chains	blue	ocean	singing	bells	blaze	murmur
0.09022	15_mother_die_gods	mother	die	gods	paris	mountain	power	hearken	harken	cloud	white
0.21823	16_mother_ring_boy	mother	ring	boy	gone	wife	child	turn'd	children	seem'd	poor
0.05924	17_fill_drink_fools	fill	drink	fools	cup	fat	conscious	empty	ancient	waiter	pint
0.07209	18_knave_kitchen_star	knave	kitchen	star	arms	mere	lead	stone	bridge	hearth	follow
0.10308	19_dress_faded_town	dress	faded	town	arms	hawk	hall	sparrow	gay	wine	eat

alpha	labels	keys (11–20)									
0.09065	0_men_woman_florian	crowd	soldier	kind	boys	girls	stir	talked	follow	grand	gained
1.79663	1_said_came_hand	good	went	men	father	child	knew	found	lay	eyes	fell
2.28222	2_love_heart_life	god	know	come	time	world	make	dark	voice	old	eyes
0.29520	3_sea_night_battle	glory	faces	thunder	death	fight	human	children	day	ship	blew
0.42260	4_answered_said_rode	turned	called	seemed	let	table	asked	saying	know	ride	looked
0.00564	5_es_wur_says	theer	mun	thaw	taäil	oän	oop	ud	thowt	upo	Roä
0.23551	6_holy_hall_saw	vision	grail	brother	quest	christ	crown	people	sware	pass	madness
0.10374	7_fool_red_neck	music	fools	moor	swine	innocence	white	marriage	run	trampled	passionate
0.14303	8_fame_charm_boon	palace	take	mood	window	careless	strange	wrought	pleasure	wise	grant
0.26351	9_spirit_light_life	lips	tears	change	brain	green	dim	thought	sense	look'd	dew
0.11458	10_brand_bold_clothed	ridge	black	wonderful	winter	bring	samite	wound	act	lightly	answer
0.08154	11_god_flesh_bread	lead	friend	saints	pillar	charge	body	foul	crowd	heaven	church
0.16291	12_ring_garden_brook	east	rhymes	brown	river	bridal	dance	spire	walks	tremble	roses
0.24222	13_earth_human_sorrow	feel	wisdom	eternal	earthly	spirit	frame	truth	nameless	noble	bloom
0.07544	14_isle_spain_sail'd	merrily	tower'd	broad	shriek	moor	spray	colour'd	paradise	slain	throne
0.09022	15_mother_die_gods	happy	golden	pine	river	vine	came	spake	dark	fairest	beautiful
0.21823	16_mother_ring_boy	years	loved	look'd	year	answer'd	ask'd	wait	kiss'd	house	grave
0.05924	17_fill_drink_fools	talk'd	head	heiress	grapes	runs	painter	pleasant	measure	theme	random
0.07209	18_knave_kitchen_star	sun	morning	knowest	life	sweetly	scorn	knaves	bow	dog	hall
0.10308	19_dress_faded_town	wheel	fall	bandit	pride	gift	host	horses	bridge	mowers	dwarf

Figure 1 shows a heatmap representation of the outputs of LDA; 20 most prominent topics are arrayed vertically, and the 66 poems are arrayed horizontally. The heatmap is

drawn by the word-weight values of the LDA runs. In the heatmap, greener cells indicate the greater density of the topic in the text in question. Conversely, more reddish cells show that the significantly lower density of the topic.

Figure 1 The heatmap of LDA topic modelling result (no. of topics: 20)



While the keys of Topic 2 include present tense verbs and nouns that deliver abstract ideas, Topic 1 is characterized by past tense verbs and concrete nouns, especially about body parts. A more careful investigation is needed, but it can be said that Topic 1 contains the narrative elements since Topic 1 tends to appear in epic poems and contains past tense verbs. 10 out of 13 poems of the *Idyls of the King* series are located as the nearest neighbours. Topic 4 appears most generally in the ten works of the *Idyls* series, and topics 6, 7, 8, and 19 are also prominent topics in the *Idyls* series. The graphological variants in past tense verbs can be seen: for example, *answered* / *answer'd*, *seemed* / *seem'd*, and *looked* / *look'd* (see Topic 4 for *–ed*; Topic 16 for *–'d* in Table 2). Tennyson primarily used *–ed* in the *Idyls* series while *–'d* appeared in non-*Idyls* poems.

The Passing of Arthur (1869) and *Morte d'Arthur* (1842) have one particular topic in common, Topic 10. Though *Morte* is not included in the *Idyls* series, these two poems were composed of a striking number of duplicate lines. The first prominent key of Topic

10, *brand*, describes a special *sword* of Arthur, called *Excalibur* in the poems (Figure 2). The second prominent key *bold* only modifies a male character *Sir Bedivere*, primarily when *Sir Bedivere* acts in response to other's utterances or acts. A plausible interpretation is that the voiced plosive bilabial /b/ is used as alliteration, and the consonant /b/ and the diphthong /oʊ/ can be interpreted as revealing *Sir Bedivere*'s grand attitudes, while the *bold* does not appear near the lines, which express Sir Bedivere's quick move with voiceless consonants.

Some words appear in multiple topics: for instance, *God*(*s*), *mother* and *ring*. *God* is a keyword of Topics 2 and 11, and *Gods* appears in Topic 15. Topic 11's keys have religious connotations, while Topic 2's keys represent general human-being aspects (Figure 2). The word *mother* appears as the primary keys of Topic 15 and 16. The difference between Topic 15 and 16 is the usage of the word. The *mother* in Topic 16 is described as 'a woman in relation to a child or children to whom she has given birth' (*OED* s.v. mother), while the mother in Topic 15 is used as a vocative with figurative meaning for a mountain (and the mountain is called *Ida* in *Œnone* (1833/1842), quoted below).

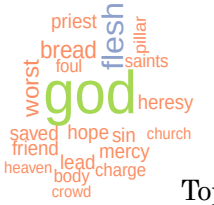
> 'O mother Ida, manyfountained Ida,
> Dear mother Ida, hearken ere I die.' (ll. 33–34)

In *Œnone*, the second line of the excerpt above is repeatedly used, and *mother* does not exactly mean a woman nor does it modify a human being, but it refers to an apostrophised *mountain* (and rhyming with *I die*). The plural *Gods* only appears in his poems about character(s) in Greek mythology or stories based on Greek mythology since Christianity adopts monotheism while Greek mythology adopts polytheism.

Figure 2 The word clouds of Topics 10, 2 and 11

Topic 10

Topic 2



Topic 11

Some words, such as, *ring* and *rose* appear in plural topics. LDA showed different part-of-speeches of the words in question belong to different topics. That is, *ring* in Topic 12 is mostly a verb, while *ring* in Topic 16 is predominantly a noun. Insight into this finding was gained by reading the concordance lines as well as a perusal of words in context. To prove that the word *ring* belongs to the different topics according to its part-of-speech,

assigning POS tags to all the words in the texts, can be one of the most applicable operations. However, both TreeTagger and Stanford CoreNLP tagger do not put tags with high enough accuracy on raw poetry texts. The biggest challenge is how to cope with poetic conventions, including the use of upper-case letters at the initial position of lines, which tends to mislead the tagger to assign wrong POS. In addition, since abbreviated words with apostrophe (i.e., o'er, howe'er, thro') are not included in standard dictionaries of the taggers, it is necessary to amend raw text poetry files (modification of line breaks and replacement of capital letters) before tagging to avoid these errors in future research.

5. Conclusion

This study has reported on the emerging results of LDA and has shown the latent topics undiscovered behind the prominent elements of the poems in the corpus. Future investigations are necessary to optimise the number of topics less arbitrarily, and to suggest how to make the best use of LDA topic modelling to analyse poetry with no less exactitude and granularity than applied to prose studies.

Bibliography

Blei, M. D., Ng, Y. A., and Jordan, I. M. (2003). "Latent Dirichlet Allocation." Journal of Machine Learning Research 3, pp. 993–1022.

Hair, S. D. (1991). Tennyson's Language. Toronto: University of Toronto Press.

Huang, C. (2020). "Quantitative Analysis of Chinese Mystery Novels: Focusing on the Works of Cheng Xiao Qing and Gui Ma Xing." Studies in Language and Culture Osaka University. 2019, pp. 31–45.

Kiyama, N. (2018). "How Have Political Interests of U.S. Presidents Changed?: A Diachronic Investigation of the State of the Union Addresses through Topic Modeling." English Corpus Studies, 25, pp. 79–99.

Meeks, E. and Weingart, B. S. (2012). "The Digital Humanities Contribution to Topic Modeling." Journal of Digital Humanities, 2, No. 1 Winter 2012, pp. 1–6.

Tabata, T. (2017). "Mapping Dickens's Novels in a Network of Words, Topics, and Texts: Topic Modelling a Corpus of Classic Fiction." Japanese Association for Digital Humanities Conference 2017, September 2017, Doshisha University.

Resources

MAchine Learning for LanguagE Toolkit: http://mallet.cs.umass.edu/index.php (Latest Accessed Date: September 6th, 2021)

Oxford Dictionary of English. 2nd ed. Revised. (2005). Oxford University Press.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 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

佐竹 由帆(青山学院大学)
t31330@aoyamagakuin.jp

The effects of DDL on learning verb-noun collocations in English

SATAKE Yoshiho (Aoyama Gakuin University)

Abstract

Although collocation knowledge is important, Japanese English learners' vocabulary learning tends to be word-based, and collocation teaching and learning are not sufficiently conducted. Since the effects of data-driven learning (DDL), which is corpus-referenced learning, on vocabulary learning have been examined in various ways, this study examined the effects of DDL on verb-noun collocation learning. The subjects were 19 Japanese university sophomores who were intermediate learners of English. Once a week for 10 weeks, they learned two verb-noun collocations by searching example sentences in the Corpus of Contemporary American English (COCA). The results of the pre- and post-tests were significantly different with a large effect size by Wilcoxon's signed-rank test (z=3.63, p=.000, r=.59), suggesting the effectiveness of DDL for collocation learning.

Keywords

data-driven learning (DDL), collocations, Corpus of Contemporary American English (COCA)

1. はじめに

コロケーションとは単語同士の恣意的かつ発生頻度の高い結びつきである (Lewis, 1997)。多くの単語はコロケーションの形で使用されるため (Sinclair, 1991), 外国語教育においてコロケーション学習は重要である (e.g., Lewis, 1997)。しかし, 日本の学校英語教育において語彙指導は新出語に偏り学習者は単語とその意味を覚えることに集中しがちであり (川, 2013), コロケーション指導・学習は十分に行われているとは言えず, コロケーションは日本の英語学習者にとって難しい項目である。コーパスを参照するデータ駆動型学習 (DDL) の語彙学習に対する効果は様々に検証されているため (e.g., Boulton & Cobb, 2017, Satake, 2020a), 本研究は DDL の動詞-名詞コロケーション学習に対する効果を検証した。

2. 先行研究

DDL とは，学習者がコーパスから得られる大量の本物のデータに触れ，自律的に言語を調べ，パターンを推測するプロセスである（Johns, 1991）。第二言語習得（SLA）の観点からは，学習者が言語データを調べて活用する DDL は，帰納的学習（Tomasello & Herron, 1988）や気づきと関連している（Chambers, 2010）。その有効性については近年様々な報告がなされており，DDL 論文のメタ分析がその有効性を示している。Boulton & Cobb (2017)は 2014 年上半期に出版された 64 本の DDL の英語論文における DDL の有効性についてメタ分析を行い，実験群と統制群，事前・事後テストの二郡間比較においていずれも効果量大程度以上であったことから，DDL は第二言語学習に大きな効果があったと結論づけている。また, Mizumoto & Chujo (2015)は 2007-2014 年に出版された中條のグループによる日本の初級英語学習者のみを対象とする 14 本の DDL 論文における DDL の有効性についてメタ分析を行い，事前・事後テストの群内比較において効果量中程度であったことから，DDL は日本の初級英語学習者の第二言語学習に中程度の効果があったと結論づけている。DDL の具体的な実践研究としては，コーパスを参照する誤り修正や（e.g., Satake, 2020a, Tono, Satake & Miura, 2014），コロケーション等の語彙学習などがあり（e.g., Flowerdew, 2010, Satake, 2015, 2020b），多様な学習項目における有効性が報告されている。

動詞-名詞コロケーションについては，学習者の使用は難しいとする報告がある。Laufer & 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，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た結果，英語学習者は英語母語話者に比べて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。

コーパス参照がコロケーションなどのパターン化されたフレーズの習得に役立つことは指摘されているが（Flowerdew, 2010），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，先行研究が少なく具体的な効果はまだ不明である。コーパス参照が日本の英語学習者の動詞-名詞コロケーション知識の向上に効果的であるかどうかを判断するには，さらなる実証的研究が必要である。ゆえに，本研究のリサーチクエスチョンは，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を記憶する上で有効か，とする。

3. リサーチデザイン

3.1 参加者

参加者は，2021 年前期に筆者が担当している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 名のデータを使用した。学生は 13 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: “I **do** a habit of taking a walk. 私は散歩するのを習慣にしています。”）が書かれたワークシートを配布され，現代アメリカ英語コーパス(COCA)を参照して15 分で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て提出した。コーパス使用を指導した初回のみ，時間の都合により英文一文，8 分で実施した。フィードバックとして，翌週対象コロケーションを使用した英文を筆者がクラス全体に提示し短い明示的解説をした。対象コロケーションは，英文法書 *English collocations in use: advanced* (O'Dell & McCarthy, 2008)に収録されている動詞-名詞コロケーションで，COCA で名詞の 3 語前までに動詞が出現する動詞-名詞の組み合わせを調べると検索結果上位 9 位以内に対象コロケーションが出現するものを 19 選び，COCA を参照して筆者が課題英文と和訳を作成した（事前事後テスト問題番号順に， 1) “make a contribution”, 2) “make a habit”, 3) “make a living”, 4) “fit the description”, 5) “stimulate growth”, 6) “play host”, 7) “broach the subject”, 8) “declare independence”, 9) “reach agreement”, 10) “propose a toast”, 11) “grant permission”, 12) “dump waste”, 13) “heal the rift”, 14) “heap praise”, 15) “make a change”, 16) “bring a halt”, 17) “provoke an outcry”, 18) “miss the point”, 19) “take exception”)。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。

3.3 参照コーパス

COCA はアメリカ英語の大規模均衡コーパスであり（Davies, 2008-），そのインターフェイスは使用が比較的簡単であることから参照コーパスとして使用した。初回のタスクの前に筆者は参加者にコーパスの使い方について 20 分指導した。筆者は名詞の 3 語前までに動詞が出現する動詞-名詞の組み合わせの検索方法とコンコーダンスラインの分析について説明し，参加者は例題（“It is a good idea to **do** a comparison between American and Japanese culture. アメリカと日本の文化を比較するのはいい考えですね。”）の不自然な動詞-名詞コロケーションの名詞を検索してより自然な動詞-名詞コロケーションの英文に修正する練習を行った。参加者のほぼ全員が例題の英文を修正することができた。

3.4 事前事後テスト

タスクの効果を検証するため，第 1 回目のタスクの前週に事前テストを，第 10 回目のタスクの 2 週間後に事後テストを行った。事前テストと事後テストは同一問題で，解答時間 5 分とした。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，“(　　) a habit 習慣にする”），タスクで学習した 19 の動詞-名詞コロケーションを出題した。正誤については，O'Dell & McCarthy (2008) 収録のコロケーションでなくても同意かつ COCA 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 19 点で採点した。

3.5 手順

本研究の手順は下記の通りである。

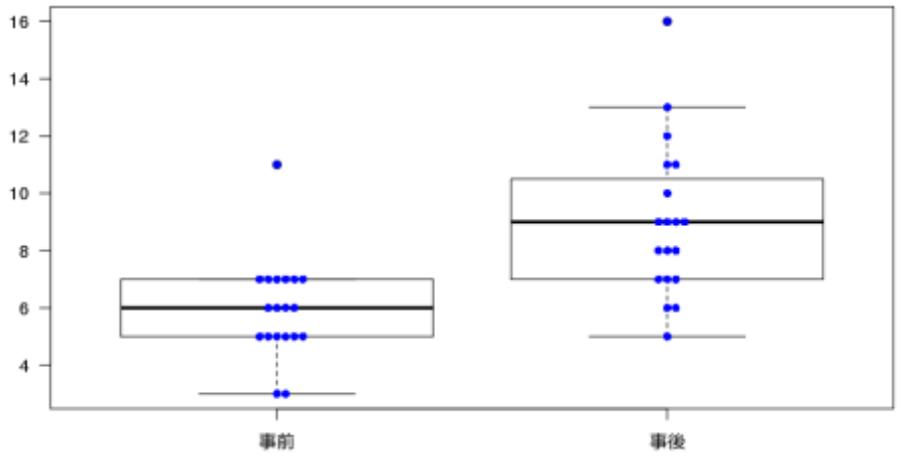
1) 事前テスト（5 分，第 1 週）
2) COCA の使用法指導（20 分，第 2 週）
3) 動詞-名詞コロケーション修正タスク 2 題（15 分，第 3-11 週，第 2 週のみ 1 題，8 分）
4) フィードバック（1-2 分，第 3-12 週，3）の修正タスク翌週）
5) 事後テスト（5 分，第 13 週）
6) 分析

タスクの有効性検証のために，事前事後テストの結果についてウィルコクソンの符号付順位和検定を行った。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。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

図 1 の事前事後テストの参加者別点数が示すように，参加者のテストの点数は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 19 点満点中 5.9 点，事後テストの平均点は 9.0 点で事後テストの方が高く，ウィルコクソンの符号付順位和検定の結果は，有意差あり，効果量大だった（z=3.63, p=.00, r=.59)。問題別に見ると，事後テストで点数が向上したのが 19 問中 15 問，同点が 3 問，低下したのが 1 問で過半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，全体的に事後テストでの点数の上昇傾向が見られたと言える。

図 1 事前事後テストの参加者別点数



一方, 問題別のウィルコクソンの符号付順位和検定の結果, 事前事後テスト間で有意に向上したのは 5 問だった(2) z=2.12, p=.03, r=.34, 5) z=2.24, p=.03, r=.36, 9) z=2.83, p=.00, r=.46, 10) z=2.00, p=.046, r=.32, 18) z=2.83, p=.00, r=.46)。表 1 の参加者 19 名全体の事前事後テストの問題別点数とタスク時の適切な修正数とタスク実施週が示すように, 上記 5 問について修正数の多少, 実施週の時期について共通の傾向は見られなかった。

事後テストの点数の向上が修正タスクの学習効果のみによるならば, タスク時修正数が事後テストの点数以上になるはずだが, 表 1 が示すように, 事後テストで点数の向上が見られた 15 問のうち, タスク時修正数が事後テストの点数以上なのは 9 問だった (4), 5), 6), 10), 13), 15), 16), 17), 18))。事後テストの点数の向上がタスク時修正の影響だけでは説明できないことから, 修正タスクに加えて模範解答を提示して説明する明示的フィードバックも事後テストの点数向上に影響したと考えられる。本研究では対照群がないため修正タスクとフィードバックそれぞれの影響の度合いについては検証することができないが, タスク翌週の 1-2 分の解答提示と説明のみに大きな学習効果があったとは考えにくいため, 修正タスクとフィードバックが相互作用して事後テストの点数向上につながったのではないだろうか。この点については対照群のある調査を今後行うことで解明したい。

以上のことから, 動詞-名詞コロケーションをコーパスを参照して修正するタスクはコロケーションを記憶する上で有効であり, 修正タスクとフィードバックの両者が寄与していると言える。

表 1 事前事後テストの問題別点数とタスク時の適切な修正数とタスク実施週(参加者 19 名全体)

問題番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
事前テスト点数	8	9	16	0	8	3	2	10	7	3	14	6	0	0	15	0	7	4	1
事後テスト点数	12	15	17	3	13	6	1	13	15	7	18	6	3	0	17	1	7	12	5
タスク時修正数	8	8	13	15	16	15	9	8	7	15	9	5	15	1	17	2	17	17	4
タスク実施週	8	10	6	3	5	2	4	9	9	3	11	10	7	7	4	5	11	8	6

5. まとめ

本研究の目的はコロケーション学習に対する DDL の効果検証であり, 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を記憶する上で有効か調査した。不自然な動詞-名詞コロケーションを COCA を参照して修正し次週に説明フィードバックを受ける学習を 10 週間行った結果, 参加者の事後テストの点数は事前テストより有意に向上した。動詞-名詞コロケーションをコーパスを参照して修正するタスクはコロケーションを記憶する上で有効であり, 修正タスクとフィードバックの両者が寄与していると考えられる。本研究はコロケーション指導及び学習の有効な方法を提案している点で, 英語語彙指導の選択肢を増やすことに貢献している。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, 動詞-名詞コロケーションの記憶にコーパスを参照する修正タスクと明示的説明フィードバックが影響する度合いについて, 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

Boulton, A. & Cobb, T. (2017). Corpus use in language learning: a meta-analysis. *Language Learning* 67(2), 348-393.

Chambers, A. (2010). What is data-driven learning? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 319-332). Routledge.

Davies, M. (2008-). The Corpus of contemporary American English. Retrieved from https://www.english-corpora.org/coca/

Flowerdew, L. (2010). Using corpora for writing instruction. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 444–457). Routledge.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing*. (*English Language Research Journal* 4, 1–16). ELR.

川貞夫 (2013).「語彙指導の諸問題と語彙学習方略の習得をめざした指導」『「英検」研究助成報告』(日本英語検定協会) 25, 186-204.

Laufer, B. & Waldman, T. (2011). Verb-noun collocation in second language writing: a corpus analysis of learners' English. *Language learning* 61(2), 647-672.

Lewis, M. (1997). *Implementing the lexical approach.* Language Teaching Publications.

Mizumoto, A. & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies* 22, 1–18.

O'Dell, F. & McCarthy, M. (2008). *English collocations in use: advanced*. Cambridge University Press.

Satake, Y. (2015). Comparison of dictionary use and corpus use: different effects on learning L2 phrases. In *Proceedings of ASIALEX 2015*, 222 – 229.

Satake, Y. (2020a). How error types affect the accuracy of L2 error correction with corpus use. *Journal of second language writing* 50.

Satake, Y. (2020b). The effects of corpus consultation on learning English collocations. *Journal of Corpus-based Lexicology Studies* 2, 13 – 30.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Tomasello, M. & Herron, C. (1989). Feedback for language transfer errors: the garden path technique. *Studies in second language acquisi*tion 11, 385-395.

Tono, Y., Satake, Y. & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL* 26(2), 147-162.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 2021

Distribution of Repeated Appearance of Grammar Items in Junior High School Textbooks through Nonlinear Regression

AMMA Kazuo (Dokkyo University)
ammakazuo@mac.com

Abstract

This paper is aimed at presenting a new smoothing method for what appears to be a disorderly distribution of frequency of text data. The target data were 25 major grammar items in the six junior high school English textbooks. Instead of simple observation of frequency per time unit, the occurrence was accumulated allowing non-linear regression analysis. The four coefficients of the cubic regression formula for each grammar item were collected as the raw data for a factor analysis intended to seek a pattern of distribution among the grammar items. The result clearly demonstrated a contrast of how the items are exposed. Some attempts of interpreting the learning/teaching process follow.

Keywords

grammar item, junior high school textbook, nonlinear regression, factor analysis

1. Introduction

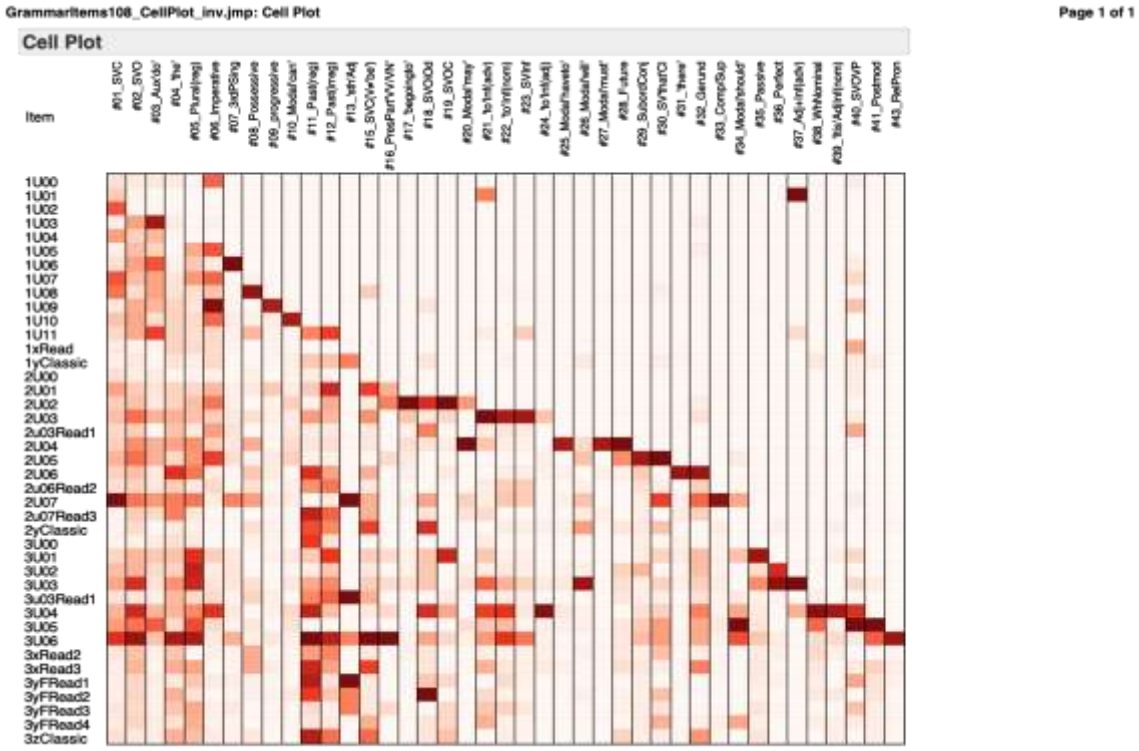
Learning at school requires repeated exposures, attempts, and evaluation — in any academic subjects but particularly language learning. However, in the Japanese junior high school (JHS) context the beginner-level grammar items are linearly arranged in the curriculum and occasional reviewing of past items seems to be neglected; once an item is introduced with a certain set of practices the item is regarded as complete. If in doubt of the mechanism and use of a particular item the learner has to refer back to the lesson where the relevant explanation is found. In all the six textbooks authorised by the Ministry of Education (MEXT) the syllabus is constructed in a Structuralist fashion — unlike the cyclic syllabus in the early days of communicative language teaching. It is this aspect of item reappearance that our strength lies in, stepping forward from counting the total frequency (cf. Hayashi, et al. 2016; Ishii 2016).

Take *New Horizon*, for example. Figure 1 indicates the density of occurrence frequency of 41 items conducted in a separate study (Amma, 2018). Within each item (horizontal scale) the relative frequency rate was converted to the darkness of the cells corresponding to lesson units (vertical scale; '1U01' means Book 1, Unit 1, for example). Whereas the first five items (#01 - 05) and two more items (#11, 12) show a relatively diffused distribution, the other items appear most frequently on introduction and are seemingly forsaken in the rest of the learning period. The *3rd Person Singular* (#07), for example, is introduced in Unit 6 of Book 1, halfway through the first year grade as the primary peak, but the next time it appears with some density is about a year later (Unit 7 of Book 2). Even worse are *Progressive* (#09), *Be Going To* (#17), *Modal* 'have to' (#25), *Modal* 'must' (#27), *Existential* 'there' (#31), and *Relative Pronoun* (#43). Although some items appear even before a formal introduction (such as 'Nice to meet you' for *To-Infinitive Following Adjective* (#37)), others are typically one-off.

Apart from the distorted distribution of reappearance the frequency of instances to reappear per unit is unpredictable and disorderly. The purpose of this study is to seek the relevance of application of a nonlinear regression, namely cubic curves. It the method proves statistically practical, it can be used to generalise the occurrence

frequency of what appears to appear at random. New to this study is the exhaustiveness of data; unlike the previous report (Amma, 2018) where *New Horizon* was the only source of texts, we have included all the texts in all six textbooks, including 53882 sentence lines and approximately 317 thousand words.

As an extension of this method we conducted a factor analysis, though in progress. We extracted coefficients of the first-, second-, and third-order terms of the cubic regression formulae as well as the intercept and used them as new independent variables. The factor analysis revealed a two-dimensional space where the grammar



items are distributed, thus enabling the generalisation of reappearance patterns.

Fig. 1: Cell plot of frequency density (from Amma, 2018)

2. Data

The present data was taken from all the junior high school English textbooks authorised by MEXT, published in 2016: *Columbus 21*, *New Crown*, *New Horizon*, *Sunshine*, *Total*, and *One World*. The database includes:

- all the text printed in the student books
- audio scripts and keys to exercises printed in the teacher's manual
- expected output, written and oral, in the exercises in the student books.

In other words, all the amount of written and spoken text input/output is contained. Where the exercise specify repetition, the corresponding text is duplicated the number of times for repetition. Where the answer is locally dependent, a generic answer was provided. If the instruction is 'Find three classmates and introduce yourselves', for example, then the pattern 'My name is NAME. Nice to meet you.' is repeated six times, where there are six instances of *BE verbs* (Sentence Type II) and six instances of *to-infinitives following an adjective*.

The raw data was sent to Wmatrix and had all the words assigned parts of speech.

Japanese proper names (persons and locations) are assigned <NP1> (proper noun singular) and Japanese words are assigned <FW> (foreign word).

Grammar tags are added to the text manually, using text editors and Excel. The following sentence (1) was taken from *Columbus 21 Book 3*, p.93 and was tagged as (2).

(1) I think boxed lunches are better because you can choose what to eat.

(2) 8353 I_<PPIS1> think_<VV0>**<V3N>** boxed_<VVN>**<VAM>** lunches_<NN2>**<NPR>** are_<VBR>**<V2B>** better_<JJR>**<CMP>** because_<CS>**<SCJ>** you_<PPY> can_<VM>**<AX1>** choose_<VVI>**<V3N>** what_<DDQ>**<WHN>** to_<TO> eat_<VVI>**<INF>** ._<.>

The list of 25 grammar items for the present and future analysis is shown in Table 1. The code number was based on the average order of formal introduction in the six textbooks, slightly different from Figure 1.

Table 1: Grammar items — codes with examples

Code	Explanation	Example
#01_V2B	'be' as copula in SVC	Here's your change. / I was surprised.
#02_PRE	preposition	I look forward to you mail.
#04_DEF	article 'the'	She likes some of the players.
#05_NPR	plural -s	How many CDs do you have?
#06_ADO	auxiliary 'do'	Do you play the piano?
#07_IMP	imperative	Excuse me. / Let's try basketball.
#09_3PS	third person singular simple present -s	Sakura has a guitar. / Does she like it?
#10_AX1	modal auxiliary 'can'	We cannot ride a bike.
#11_PRG	present progressive -ing	Kevin is playing tennis now.
#14_VGM	'go' V-ing	I went shopping with a shopping bag.
#15_GNG	future (be going)	How long are you going to stay?
#16_V2C	non-be copula in SVC	You look happy.
#17_AX0	future (will/shall)	I will show you some pictures.
#19_SCJ	subordinate (when, if, because, etc.)	If you are interested, please call us.
#20_AX2	deontic modal auxiliary 'may'	May I ask your favor?
#21_AX3	deontic modal auxiliary 'must'	You must wash in the bathtub.
#22_AX4	deontic modal auxiliary 'have to'	Do I have to eat everything?
#24_EXS	existential 'there'	There are many cherry trees in the park.
#26_GRD	gerund	Making plastic bags causes global warming.
#27_INF	to-infinitive (inclusive)	She was made to appear.
#30_AX5	deontic modal auxiliary 'shall/should'	Shall I take your picture?
#33_PSV	passive voice	Tickets are sold online.
#34_PRF	perfective aspect	How long have you lived here?
#38_VPM	postmodification with participle	This is a food made from cacao beans.
#39_REL	relative pronoun	This is the message he left for us.

3. Analysis

For each sentence its location was calculated as the relative position in the entire set of sentences within the textbook. Further, the point of formal introduction was set to 0 and the end point (ie., graduation) was set to 1.

The occurrence frequency was accumulated each time an instance was observed. The cumulative data was normalised so the value would take between 0 and 1 (which is the end point).

These conversions were introduced to enable the comparison of grammar items of

different introduction points and different magnitudes of total frequency.

The normalised data was put to analysis by a statistical tool JMP (JMP 2021). For the entire set of textbooks and individual textbooks a cubic regression expression was applied. Cubic regression was selected instead of quadratic regression because the overall squared residuals indicated a 3.3% improvement for the former (Amma, 2018).

A cubic regression was applied to each item. When the regression formula

$$y = ax^3 + bx^2 + cx + d \quad (3)$$

was obtained, the four coefficients *a*, *b*, *c*, and *d* were turned to a new set of variables characterising the cumulative distribution of the item. 25 sets of coefficient values were collected as the new data for factor analysis.

4. Results

4.1 Regression

Table 2 summarises the squared residuals with observations (total frequencies) after cubic regression was applied to each item.

Table 2: Squared residuals as a result of application of cubic regression (sorted in the descending order of RSquare)

Code	RSquare adjusted	Observations
#02_PRE	0.997004	52606
#01_V2B	0.993585	52715
#04_DEF	0.99182	51720
#05_NPR	0.989752	51141
#07_IMP	0.976432	49010
#06_ADO	0.975956	51294
#09_3PS	0.975693	46724
#17_AX0	0.960601	43360
#27_INF	0.9592	27477
#10_AX1	0.956443	43360
#11_PRG	0.955507	41377
#19_SCJ	0.942701	43049
#34_PRF	0.934153	15583
#39_REL	0.931075	7758
#26_GRD	0.910087	27243
#33_PSV	0.90934	19601
#15_GNG	0.862849	33286
#24_EXS	0.837404	29732
#16_V2C	0.809846	37265
#38_VPM	0.794032	8558
#20_AX2	0.74826	30259
#21_AX3	0.72966	30912
#22_AX4	0.629177	29418
#14_VGM	0.622372	37354
#30_AX5	0.5292	26471

Most of the items were fairly well summarised with high squared residuals. 16 items out of 25 scored more than 0.9. Figure 2 shows the item with the best score, *Preposition* (#02).

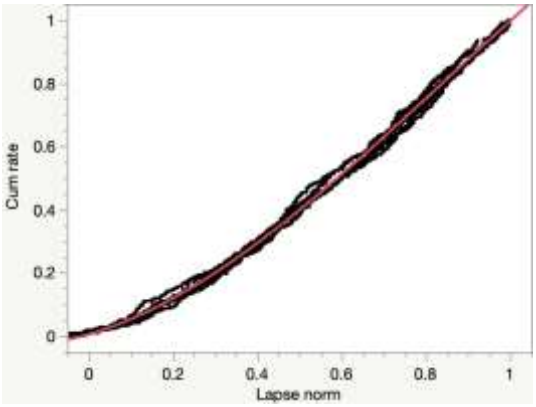


Fig. 2: A cubic regression applied to *Preposition* (#02)

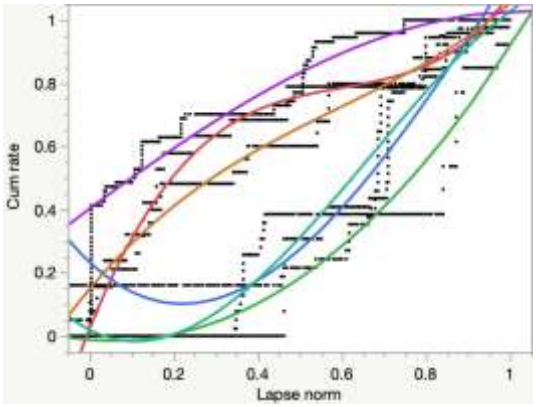


Fig. 3: A cubic regression applied to *Auxiliary verb* 'should' (#30)

In contrast, the item with the worst squared residual score was *Auxiliary verb* 'should' (#30), as shown in Figure 3. The coloured curves represent individual textbooks. Judging roughly from the distribution of items in Table 2, high-frequency items with young code numbers tend to be stable with little variability whereas auxiliaries with codes AX*n* (except for 'will'/'shall' (#17_AX0)) and *post-verb modifying participles* (#38_VPM and #14_VGM) resulted in a wide variability across textbooks.

4.2 Factor analysis

Using the coefficients of the cubic regression formulae a factor analysis was conducted (Maximum likelihood, Varimax). Individual items were assigned factor scores and plotted in a two-dimensional space (Fig. 4).

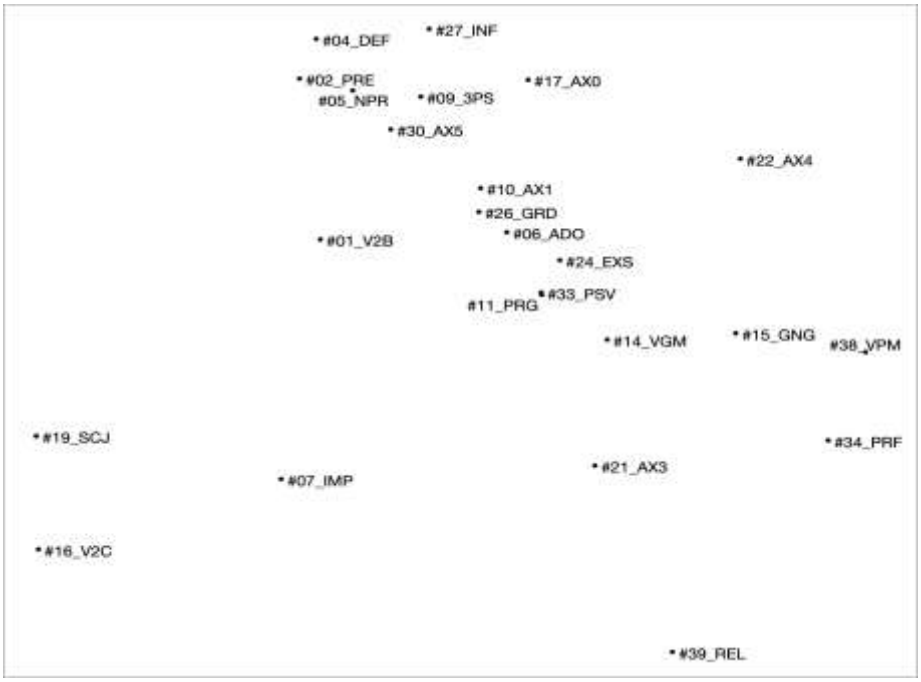


Fig. 4: Two-dimensional distribution of items after factor analysis

5. Discussions

Figure 4 indicates a pattern of distribution. On top left are mostly young numbered items with high frequency. These are the grammar items that appear in a variety of genres with a common regression curves in *concave* shapes, as represented by Figure 2. Towards bottom right are items in relatively old numbers, displaying regression curves in *convex* shapes, as represented by Figure 5. The contrast in shapes suggests how the items reappear. High-frequency items, because of their universal nature, tend to appear more as the learning stage develops with more opportunities of exposure. In contrast the items that appear in late stages do not have sufficient opportunities of exposure in the rest of the school period. As a result, the items are introduced with some intensity followed by a short span of reviews and exercises. It is items of this pattern that have a potential risk of neglect unless sufficient follow-ups are provided. High-frequency items are structurally simple and have relatively few problems. There will be an increasing number of exposures as time passes, where the learnt rules are applied to comprehension and production, thus this success experience is enhanced, hopefully. However, should one fail to use it properly in the early stage the crack might enlarge as the exposure is accelerated.

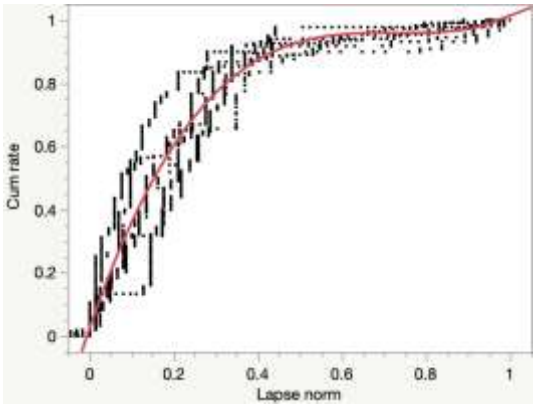


Fig. 5: A cubic regression applied to Relative Pronoun (#39)

The variability of distribution in some items seems to suggest that these items are topic-dependent. Deontic modals are usually directed from the speaker to the addressee in the discourse. The use of the item may reflect more directly the discourse plot the textbook writer has in mind than semantically neutral high-frequency words. Another possibility is the false recognition of the zero point. In the case of 'should' (#30) three textbooks show a low-rise pattern with a long initial dormant span (Fig. 3). In fact three of the six textbooks do not present the item with a formal introduction. If the true starting point is recognised properly the distribution patterns will converge.

This survey is time-consuming especially due to the correction of errors in text recognition and POS assignment. Although the textbook revision is faster than the pace of the survey, the statistical procedure should remain valid across all versions. Further attempts, including addition of grammar items, are sought for increased precision and validity of accountability.

References

AMMA Kazuo. (2018). Extracting patterns from transition of occurrence frequency of grammar items in a junior high school textbook. *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference* (APCLC 2018), 219-226.

[Hayashi, et al.] 林正頼・石井康毅・高村大也・奥村学・投野由紀夫. (2016).「CEFR-based Coursebook Corpus からの CEFR レベル別基準特性の特定」. 投野由紀夫(代表)『学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究』(平成 24 年度～平成 27 年度科学研究費補助金(基板研究(A))研究課題番号 24242017 研究成果報告書).

[Ishii] 石井康毅. (2016).「CEFR-J Grammar Profile の構築のための英文法項目の選定・抽出・頻度集計・精度評価」. 投野由紀夫(代表)『学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究』(平成 24 年度～平成 27 年度科学研究費補助金(基板研究(A))研究課題番号 24242017 研究成果報告書).

JMP. (2021). *JMP Pro* Version 16.0.0. Cary, NC: SAS Institute.

『英語コーパス学会大会予稿集 2021』（ISSN　2436-6447）
Proceedings of the JAECS Conference 2021

Classroom Application of a Web-based DDL Support Tool in a Secondary School

NISHIGAKI Chikako (Chiba University)
AKASEGAWA　Shiro (Lago Institute of Language)
KAWANA Takayuki (Junior High School Attached to Chiba University)
NAKAI Kohei (Junior High School Attached to Chiba University)
KENMOKU Shinya (Junior High School Attached to Chiba University)
YAMAZAKI Tatsuya (Junior High School Attached to Chiba University)
gaki@faculty.chiba-u.jp, lagoinst@gmail.com, kawana@chiba-u.jp,
nakai0526@chiba-u.jp, kenshin@chiba-u.jp, yamazaki18@chiba-u.jp

Abstract

Although DDL is becoming more widely used globally, there are very few studies conducted on introductory-level students. This is due in part to the lack of suitable corpora and user-friendly search software. We developed a teaching-oriented corpus and a web-based DDL tool, called hDDL, to address this lack, and in this paper, we report its development and the results of a pilot study. Two groups of Japanese 7th graders (N=139) learned SVOO structures using hDDL. Both groups were given an error correction task; only the treatment group was given a noticing guide. Both groups improved their scores on the error correction test. However, the number of occurrences of noticing the SVOO sentence structure on the worksheets was statistically higher in the treatment group.

Keywords

data-driven learning, teaching-oriented corpus, introductory-level, secondary school

1. Introduction

1.1 Background of the study

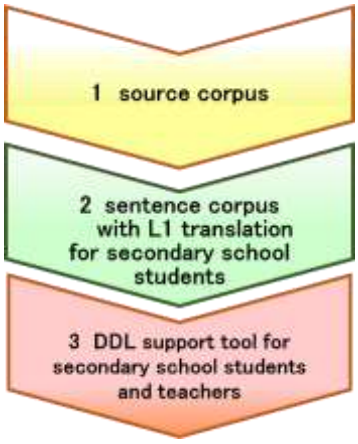
Although classroom applications of DDL are expanding worldwide, Wicher（2020）points out that DDL usage with young learners is not widespread. He lists three reasons for this: (a) a lack of appropriate student-friendly corpus resources, (b) teachers' lack of awareness of corpus applications, and (c) the lack of empirical studies testing the actual effects of DDL. Concerning the lack of resources, one notable exception is called the Sentence Corpus of Remedial English (SCoRE: https://www.score-corpus.org/). Although SCoRE was created for lower proficiency-level university students, much of its ideology and structure was used as a basis for creating a similar corpus and tool for secondary

school students called hDDL. This study was conducted to address the lack of appropriate corpora for young learners by (a) developing a teaching-oriented corpus that matches the English proficiency level, developmental stage, and interests of secondary school students (junior and senior high school students in Japan), and (b) developing a web-based DDL tool, called hDDL (https://h.ddl-study.org/), that is used to search the corpus. Then, (c) we used hDDL in an English class of 7th graders (first grade of junior high school in Japan) to examine the effect of DDL on these introductory-level young learners.

1.2 Development of hDDL

The process for creating hDDL is shown in Figure 1. First, we developed a "source corpus" consisting of approximately 24 million words. The data were collected from government-authorized school textbooks published in Japan, China, Korea, and Taiwan, reading textbooks used in U.S, graded readers used for ESL extensive reading, and ESL material found on the Internet.

Figure 1 *Process to Create hDDL*



Second, native English speakers and Japanese English teachers collaborated to create a sentence corpus by referring to this source corpus. Each English sentence appears with its L1 (Japanese) translation. The corpus currently consists of about 12,500 words and 2,350 sentences, and new sentences will be continually added.

Third, we created a web-based DDL tool called hDDL for secondary school students to learn vocabulary and grammar. hDDL has three sub-tools: (a) a user-friendly concordancer which displays search results either as regular sentences or as keyword in context (KWIC) and allows for sampling and sorting (Figure 2); (b) a simple pattern browser so students can view sentences by grammar item (Figure 3); and, (c) a fun word arrangement quiz that automatically makes and marks word arrangement quizzes. It was created to motivate students and detect their weaknesses in grammar. hDDL also has additional special functions for its target learners. The students can listen to the pronunciation of the searched sentences; they can search for English sentences by specifying the number of words included in sentences, for example, three words (e.g., "I enjoy camping") or four (e.g., "I enjoy doing housework").

1.3 Objectives and Research Questions

This study introduced hDDL with 7th graders whose English grammatical

Figure 2 *Concordancer of hDDL*

Figure 3 *Pattern browser of hDDL*

knowledge at an introductory level. The students had almost no experience with DDL. In addition, the students' base of grammatical knowledge was not sufficient to discover the grammar rules on their own. Therefore, two different types of scaffolding were given to elicit students' noticing and allow them to learn autonomously and inductively. We examined the differences in the acquisition of grammatical knowledge depending on the scaffolding.

RQ1 Can hDDL be used to lead 7th graders to inductive grammar learning?

RQ2 What differences in grammatical knowledge acquisition does different scaffolding elicit from 7th graders?

2. Methodologies

2.1 Participants

Four classes with 139 7th graders who were affiliated with a national university participated in this study. Seventy-one students were in the treatment group, 68 students were in the control group. They met for one 50-minute DDL lesson. All students began learning English in elementary school from the fifth grade at the latest, while some had begun from the first grade. Explicit grammar instruction started in the 7th grade. This study was carried out in March 2021, at the end of their academic year.

2.2 Learning targets

The learning objective was for students to acquire the grammatical knowledge of the SVOO structure of a verb *show*, and to notice and correct its incorrect usage in a sentence. Students will learn the SVOO structure in the 8th grade. The target word was *show*. Since teachers use *show* in the classroom (e.g., *show me your answer*), it can be assumed that the students had a hypothesis about the usage of *show*. Therefore, it would be interesting and informative to see if, as a result of using hDDL activities, students would be able to transfer their implicit knowledge of *show* to explicit knowledge. This change in knowledge corresponds to the process of transforming "input" into "intake" in the second language acquisition model.

2.3 Procedure

The DDL performed by students included individual, pair, and group work. Prior to this study, students had had a session to familiarize themselves with hDDL. All students were given a supplemental worksheet for the DDL class that contained an error correction task (see Figure 4). All sentences on the error correction task worksheet contained the word *show.* The sentences on the hDDL screen were not identical to those on the worksheet, but were similar in sentence length and vocabulary level.

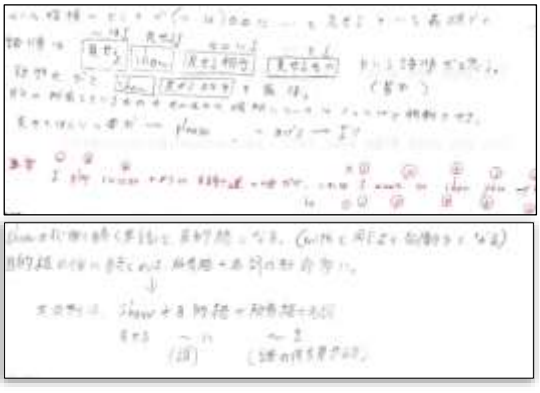
In addition to the error correction task, the teacher provided the treatment group with "noticing guides" to how and what they could find in the hDDL concordance lines. The given guides included three questions (in Japanese): (1) Looking at the first word to the right of *show,* what kind of word or words are common?; (2) What is shared among the word or word groups (e.g., *the book*, *her sister*) that come second to the right of *show*?; and, (3) What word order do you find in the sentences? The treatment group completed the error correction tasks with the guides, while the control group completed the error correction task without the guide.

2.3.1 Individual Work

First, all students were directed to an hDDL "start" webpage. They searched for *show* by entering this word in the search bar. They were asked to observe the sentences and compare the sentence on the worksheet with those on the hDDL screen. If there was an error on the worksheet, they corrected it. They were next asked to write down any rules that they discovered about *show* on the worksheet. Figure 5 shows examples of students' notes.

Figure 4 *Error Correction Task*

Figure 5 *Students' Notes*



2.3.2 Pair Work

Next, students compared their answers for the error correction task with a partner and exchanged opinions about the findings that they drew from the activity. Pair work was conducted before group work because Japanese students are shy. They exchanged opinions in small groups to get a sense of security before moving on to group work.

2.3.3 Group Work

After the pair work, the teacher and students reviewed the answers together. The teacher also elicited opinions on the sentence structure of *show* and other findings from the activity, and shared these as a group discussion in class.

2.4 Measurements

Pre- and post-tests were used to measure learning outcomes for the ability to identify and correct errors correctly. The format of the test was the same as the error correction task. The task was to find errors in sentences and change them to the correct forms. The post-test was administered in the next English class. The pre-test and post-test were identical. In addition, the students wrote down their findings from observing the concordance lines. These notes were collected and analyzed.

3. Results and Discussion

3.1 Pre-test and Post-test

Table 1 shows the results of the pre-tests and post-tests. The Cronbach's *alpha* was .645. A two-way ANOVA showed that the score increases between pre-tests and post-tests were statistically significant. (F (1,137) = 142.44, p < .000, *generalized* η^2 = .50). The effect size was "large." However, the difference between the student groups was not statistically significant (F (1,137) = .273, p <.602, *generalized* η^2 = .00). Nor was there any interaction effect between tests and student groups (F (1,137) = 2.88, p <.092, *generalized* η^2 = .01). From this, we can understand that the use of the hDDL activity in the classroom was effective for the acquisition of the grammar knowledge of this new SVOO structure. Regarding the first research question, since both groups increased the test score statistically, we can conclude that hDDL can be used to lead 7th graders to inductive grammar learning. However, we still do not know to what extent hDDL works without error correction tasks.

Table 1. *Means, Standard Deviations, and 95% CI of Pre and Post Test (Full mark=5)*

Group	*N*	Pre-test		Post-test	
		M (SD)	95% CI	*M (SD)*	95% CI
Treatment with guide	71	2.10 (1.51)	[1.75, 2.45]	3.70 (1.09)	[3.41, 4.00]
Control without guide	68	2.19 (1.49)	[1.83, 2.55]	3.40 (1.45)	[3.09, 3.70]

3.2 Students' Worksheet Notes

Students' notes on the worksheets were examined, and the number of students who

described the SVOO structure or its word order was counted (Table 2). Fisher's Exact Test showed that more students in the treatment group (with the noticing guides) found and described the SVOO structure explicitly than the control group (without the guide) (p =. 000). Since the treatment group with the guide increased the test score significantly higher than that of the control group without the guide, for RQ2, it can be said that a direct guide that indicates how to analyze the sentence structure will help introductory-level learners find grammar rules. Thus, we can see that some scaffolding is needed to elicit inductive learning like DDL for introductory-level students.

Table 2. The Result of Fisher's Exact Test

Group	With Memo	Without Memo
Treatment (with guides)	44	27
Control (without guides)	18	50

4. Conclusion

There are few examples of DDL applied to young learners in Japan and abroad. This study shows that DDL could be effective for young learners using hDDL with error corrections tasks and guides. In the study, error correction task and guides for eliciting findings were given to the students because their English level is not proficient, and they do not know how to interpret the concordance lines on the screen. Through hDDL with the error correction task, students were able to acquire grammatical knowledge to help them monitor errors in English. It was also found that they were able to elicit noticing of sentence structures if they were given an appropriate guide. It turns out that with the right DDL tools and the proper guides, teachers can implement DDL to secondary school students who begin to learn English grammar in a formal setting. In this study, we focused on a single grammar item, but in the future, we would like to conduct DDL over a more extended period to verify its effectiveness. We would also like to examine the effects of DDL on a variety of grammar items.

Acknowledgments

This research was supported by JSPS KAKENHI Grant B Number JP16H03441 and Grant B Number JP20H01277

Reference

Wicher, O. (2020). Data-driven learning in the secondary classroom: A critical evaluation from the perspective of foreign language didactics. In P. Crosthwaite (Ed.), *Data-Driven Learning for the Next Generation* (pp. 31-46). Routledge.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 2021

動詞の意味はトピックから推測できるのか
―英語の動詞 *run* を例に―

木山 直毅（北九州市立大学）／ 渋谷 良方（金沢大学）
n-kiyama@kitakyu-u.ac.jp / y.shibuya@staff.kanazawa-u.ac.jp

Exploring the meaning of 'run' with topic models

KIYAMA Naoki (The University of Kitakyushu) /
SHIBUYA Yoshikata (Kanazawa University)

Abstract

How do people understand polysemy? Various studies have addressed this question. Among them, corpus studies based on cognitive linguistic theories have argued that, in addition to the collocation of words, a great number of factors, such as the morphosyntactic environment in which words occur and their semantic types, can affect the understanding of their polysemy. An area that has yet to be fully explored is the relationship between topic and polysemy. In this study, we discuss word polysemy using a topic model. Here, we report the results of analyzing the polysemy of the English verb *run* using the biterm topic model (Yan et al. 2014). The data was extracted from the NOW corpus (Davies 2016). The results suggest that the topic on which *run* is used has an effect on its understanding. This implies that there may be an inseparable relationship between the meaning of a word and the topic on which it is used. Based on the results of this case study, we argue that it is useful to conduct research from the perspective of topic in polysemy research.

Keywords
多義語, トピック, biterm topic model

1. はじめに

人は多義語をどのように理解しているのか。この問題に取り組むべく，言語学ではこれまで様々な研究がなされてきた。その 1 つが，コーパスから得られた事例を用いてコロケーションを抽出する研究である（詳細は Hunston 2002 等を参照）。このアプローチよりも統計的に洗練された方法論を用いたのが Stefan Gries の研究に代表される量的コーパス言語学のアプローチであった。そこでは，多義語の意味は，コロケーションに加え，語が現れる統語情報（構文，句，節），アスペクトやムードに関わる形態素などの様々な要因が関わることが論じられている（e.g. Berez and

Gries 2008; Divjak and Gries 2006; Gries 2006, 2010）。本研究では量的コーパス言語学の流れを汲みつつも，語が現れるトピックも多義語の意味を決める 1 つの要因となることを論じる。

本稿の構成は次のとおりである。まず 2 節では本稿が立脚する理論的背景を概観する。その後，本稿の仮説を導入し，その仮説を検証するために用いたコーパスや分析手法を紹介する。第 4 節では調査結果を報告し仮説を実証する。最後に本稿のまとめと今後の課題を示す。

2. 先行研究

構造主義言語学や形式主義言語学（生成文法など）では，語の意味を言語内的観点から規定することが多い。これとは対照的に，百科事典的意味論では，話者が語あるいは概念に対して持つ知識との関連で語の意味を規定する（Croft and Cruse 2004; Langacker 1987）。百科事典的意味論では，例えば，英語母語話者は *photograph* という単語に対して，少なくとも (i) 風景や人の描写，(ii) 典型的には紙に現像され，その他の媒体で保管される，といった知識を持つと考えられる（Taylor 2002，一部抜粋）。これらの知識との関係を示す事例として，Taylor (2002) は下記のものを挙げる。

(1) a. The photograph is torn.
b. I'll send you the photograph as an electronic attachment.

(1a) では，写真が紙に現像されるものであることを知らずに「写真を破る」という行為を理解することはできない。(1b) では，写真の保存方法を知る必要があり，ここでは (ii) が前景化される。

(1) に見られるように，百科事典的意味論では，他の枠組みでは語用論的意味として扱われてきたものも語の意味の一部として認め，話者が語あるいは概念に対して持つ知識全体との関係に基づき，語の意味が規定される（Croft 1993: 336f）。

3. リサーチデザイン

3.1 研究目的と研究設問

上述の Gries らによる一連の研究では，語の意味を分類する上で形態素，統語，共起語の意味タイプ，他動性といった数多くの要素が関わることが示された。Gries らの研究は多義性の研究に対して重要な知見を提供するものであることは疑いがないが，本稿ではもう一つ別の視点から多義性の問題を論じたい。それはすなわちトピックの観点から論じることである。

意味が文脈依存であることは広く受け入れられている。しかし，文脈には様々なものが含まれている。本発表では，文脈の一部を構成するであろうトピックの観点から多義性を論じたい。例えば日本語の「汗を流す」の比喩的意味には，少なくとも (i)「一生懸命に働く」と (ii)「風呂に入る」の 2 つの意味がある。災害救助のトピックで用いられる場合には (i) の意味が想起されるだろうし，温泉宿のトピックで用いられる場合には (ii) の意味が想起されるであろう。これらの簡単な事例の中にも，語の意味とトピックの関係を見出すことができる。

3.2 データ

本研究では，News on the Web Corpus (Davies 2016, 以下，NOW コーパス) を用いた。同コーパスには，2010 年以降にオンライン上で出版された英字ニュース記事が収められており，2021 年 9 月 5 日の時点で 133.5 億語が収録されている。本研究では NOW コーパスに含まれる 2013 年の米国サブコーパスを利用した。これだけでも公式発表によると約 8,500 万語が収録されており，British National Corpus の Written サブコーパスとほぼ同等のサイズである。本研究では，このサブコーパスより英語の動詞 *run* とその活用形の事例を抽出し，トピックモデルの手法を用いて分析した（動詞 *run* の分析は Gries 2006; Langacker 1988; Tuggy 1988 参照）。

3.3 手法

3.3.1 トピックモデル

トピックモデルとは，文書を分類する手法である。「トピック」とは，日本語の「話題」に概ね相当する。たとえば，「この地域はモツ鍋が有名だが，活きイカや刺し身はもっと美味い。」という文を読むと，多くの人は「食事」に関する文だと理解するだろう。人は文書が何に関して書かれたものであるかを瞬時に判断する能力を持つが，そのような文書分類を統計的に行うのがトピックモデルである。

トピックモデルにおけるトピックは語の分布によって表される（Blei et al 2003; Yan et al. 2014）。先の例において，「食事」のトピックだと理解できたのは「モツ鍋」や「刺し身」などの手がかりのおかげであろう。トピックモデルは，トピック全体（*k*）にどのような語（*w*）がどれほど現れるかを確率（P(w|k)）で表す。本稿で使用する手法では，この値を ø 値と呼ぶ。

今日のトピックモデルでは，潜在的ディリクレ配分法（latent Dirichlet allocation, LDA）(Blei et al. 2003) が主流だが，LDA は入力データが短い場合，トピックの抽出精度が下がる傾向にある（Yan et al. 2013）。そこで本研究では， Yan et al. (2013) が提案した biterm topic model (BTM) を用いて調査を行った（BTM と LDA の質的比較は木山 2020 参照）。

3.3.2 分析の手順

本研究では 5 つの手続きを踏み，BTM を用いて解析した（詳細は木山（to appear）と Kiyama and Shibuya 2021 参照）。まず，3.2 節で述べたコーパスより，英語の動詞 *run* とその活用形の事例を収集（L/R 10 語）した。その後，重複データを除外し，ストップワードとターゲット語を除外した。その後，5 語未満となったデータを除外し，トピック数を 10 に設定し BTM を用いて計算を行なった。なお，本研究では NOW コーパスに付与されている POS タグを利用したが，本稿で事例を示す際は視認性を高めるためにタグは除外する。また，例文を表示する際には，文脈の理解のしやすさを考慮し，分析で使用したコンコーダンスラインより長めに表示を行なっている。

4. 結果と考察

4.1 BTM を用いた結果

前節で述べた手続きから得られた結果が表 1 である。

表 1 トピック構成語, トピックと run の意味

トピック	トピック構成語 (ø 値) の上位 10 語	トピック名	run の意味
Topic 1	people, time business, new company, way, day, world, really, system	ビジネス	経営する
Topic 2	water, power, gas, energy, fuel, natural, oil, electricity, plants, battery	エネルギー	機械を動かす
Topic 3	police, away, man, house, people, toward, away, ran, saw, old	犯罪	逃走する
Topic 4	run, running, marathon, miles, run, race, shoes, ran, running, runners	長距離走	走る
Topic 5	president, office, campaign, senate, state, election, party, governor, former, candidates	選挙	出馬する
Topic 6	home, home, game, hit, season, league, runs, baseball, games, home	野球	走る
Topic 7	school, schools, business, program, health, organization, medical, government, company, news	組織	組織を経営する
Topic 8	government, money, percent, fund, million, dollars, security, deficit, billion, federal	政府と資金	使い切る
Topic 9	back, ball, game, yards, back, quarterback, football, field, team, players	アメフト	走る
Topic 10	software, android, system, applications, devices, version, operating, server, data, computer, device	IT	システムを動かす

以下では, 紙幅の都合上, 一部のトピックのみを見ていく。Topic 4 は, その構成語から「長距離走」トピックであることがわかる。本トピックで現れる *run* の具体例を見ると,「歩行者の高速移動 (fast pedestrian motion)」(Gries 2006) の意味で使われている。

(2) I'm training currently for a marathon, and I'm taking it pretty seriously, running sixty miles per week.

Topic 3 は *police* や *away*, *man*, そしてそれらを用いた具体事例などを見ると「犯罪」トピックであることがわかる。このような文脈の場合, (3) のように *run* は「逃走する」の意味であろう。

(3) ... there's the man whose picture you've seen countless times on TV over the past few days running toward your house as he returns fire to police officers in pursuit.

Topic 2 を構成する語の上位にはエネルギーに関するものが多い。トピックを構成する語を含み, かつエネルギー関連の例文には次のようなものがある。

(4) Ordinarily water power would run the machines, but when the water wheel would not run on its own a steam engine would lift water from the stream …

(4) は水力発電が機械を稼働させることを述べており, *run* は「機械を動かす」の意味だとわかる。

以上のように, 語の意味と語が現れるトピックの間には一定の関係が見られる。また, 物理的な移動を表す「歩行者の高速移動」と「逃走する」が区別されていることから, *run* の細かなニュアンスの差を BTM では (一定レベルで) 捉えられていると思われる。

4.2 意味間の関係

前節では，トピックが *run* の意味と関連していることを述べた。それでは，意味同士の間にはどのような関係が見られるのだろうか。各トピックが *run* の意味とおおよそ対応していることから，ø 値の位置関係を見ることで意味間の関係が見えてくる。そこで本稿では ø 値を Jensen Shannon Divergence と主成分分析で次元圧縮した。その結果が図 1 である。

図 1 トピック（run の意味）間の関係

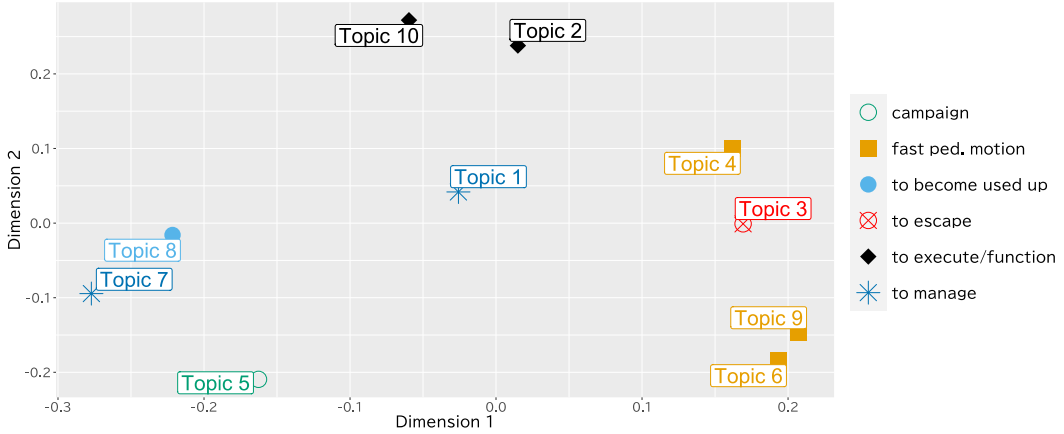


図 1 の右側（0.15 から 0.2 の間）にはスポーツに関するトピックがまとまり，「歩行者の高速移動」を表している。また，そのまとまりには Topic 3 の「逃走する」が含まれる。図の上には，テクノロジーやエネルギーに関するトピックが位置し，「機械を動かす」の意味でまとまっている。左下には「選挙に出馬する」や「(組織・会社を) 経営する」の意味がまとまっている。この配置より，右側には「文字通りの動き」の意味を，x 軸上の 0（厳密には 0.015）より左側には「複雑な構造のものを動作させること（to cause to complex entities to work)」という抽象的な意味を抽出することができる。

5. まとめと結論

本稿では，多義語の意味をトピックから推測することができるのかという問いのもと，英語の動詞 *run* を用いて調査した。BTM では *run* の多義性が概ね捉えられた。また，BTM の分析結果を図式化することにより，*run* が持つ 2 つの大まかな意味クラスを抽出することができた。

もちろん，全ての語彙の意味がトピックと関連するとは限らず，たとえば，冠詞や前置詞，あるいは一部の動詞（使役動詞の *make*など）はトピックの影響を受けにくいことが予想される。今後は，トピックの影響を受けやすいものと受けにくいものの違いを調査する必要がある。

謝辞

本研究は第 21 回日本認知言語学会にて松本曜氏（国立国語研究所）からいただいたコメントを出発点としている。記して感謝申し上げる。本研究は JSPS 科研費 20K00667 の助成を受けた。

引用文献

Berez, A. L., & Gries, S. T. (2008). In defense of corpus-based methods: A behavioral profile analysis of polysemous *get* in English. *Proceedings of the 24th Northwest Linguistics Conference*, *27*, 156–166.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Croft, W. (1993). The role of domains in the interpretation of metaphors and metonymies. *Cognitive Linguistics*, *4*(5), 335–370.

Croft, W., & Cruse, A. (2004). *Cognitive linguistics*. Cambridge University Press. Cambridge.

Davies, M. (2016). Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. https://www.english-corpora.org/now/

Divjak, D., & Gries, S. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory 2*(1), 23–60.

Gries, S. T. (2006). Corpus-based methods and cognitive semantics: The many senses of *to run*. In S. Gries & A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics corpus-based approaches to syntax and lexis* (pp. 57–99). Mouton de Gruyter.

Gries, S. T. (2010). Behavioral profiles. *The Mental Lexicon*, *5*(3), 323–346.

木山直毅. (2020).「アメリカの新聞が喚起する man と woman の知識差」.『統計数理研究所共同研究レポート』, 432 巻, 49–63.

木山直毅. (to appear).「意味論・語用論とコーパスのインターフェイス」. 米倉よう子 (編著)『意味論・語用論と他の分野とのインターフェイス』. 開拓社.

Kiyama, N., & Shibuya, Y. (2021). Applying topic models to study polysemy: The case of the noun *streams*. *Papers from the 21st National Conference of the Japanese Cognitive Linguistics Association*, *21*, 291–303.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites (vol. 1)*. Stanford University Press.

Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (pp. 127–161). John Benjamins Publishing Company.

Taylor, J. R. (2002). *Cognitive grammar*. Oxford University Press.

Tuggy, D. (1988). Náhuatl causative/applicatives in cognitive grammar. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (pp. 587–618). John Benjamins Publishing Company.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web*, 1445–1456.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 2021

現代スペイン語における主語後置の数理モデル化

小林 純一朗(東京外国語大学 学部生)
kobayashi.junichiro.t0@tufs.ac.jp

佐野 洋(東京外国語大学)
sano@tufs.ac.jp

A Study on a Mathematical Modeling of Subject Postposition in Contemporary Spanish

KOBAYASHI Junichiro (Tokyo University of Foreign Studies, Students)
SANO Hiroshi (Tokyo University of Foreign Studies)

Abstract

The study aims to build a logistic regression model which explains/predicts the occurrence of subject postposition in intransitive sentences in contemporary Spanish. Previous statistical researches, such as Brunetti & Bott (2011), employed variables like 'generic' feature of the subject, based on a traditional idea that information structure is a critical factor (Contreras 1976). However, given our corpus size (two billion words), this scrupulous methodology was not a realistic option for our large-scale study. Because of this, we utilized 25 variables, all of which are formalistic (e.g., 'definiteness' of the article). Our model from over 20,000 samples was generally better than previous one (Brunetti & Bott 2011: 28). Also, coefficient values might be a clue to the correspondence between surface variables and the information structure, which is of cross-linguistic interest.

Keywords

スペイン語, 主語後置, 情報構造, 表層特徴, ロジスティック回帰

1. はじめに

現代スペイン語(以下, スペイン語)の自動詞構文では, しばしば述語動詞(時制に応じて語形変化した定形動詞)に対して主語(述語動詞と人称・数の一致がある名詞句)が後置される。後置の要因分析において, 先行研究は定性分析から仕組みを探る機能主義的概念(情報構造)に拠ることが多い。対して本研究は, 定量分析・計量主義的アプローチを試みた。具体的には, 冠詞の定・不定や時制などの表層から得られる形態統語論的な特徴から主語位置を説明する数理モデルの構築を目指した。本研究では, 予測モデルとしてロジスティック回帰分類器を採用する。この分類器のパラメーター(偏回帰係数など)を人が言語学視点で解釈することで, スペイン語における主語後置のメカニズムを明らかにしようとする試みである。

2. 先行研究

2.1. 自動詞構文における主語後置

寺崎(1982: 106-109)によれば，スペイン語で主語後置が頻繁に起こるのは，(1)提示型動詞，(2)感情経験を表す動詞，(3)再帰構文，(4)判断文の 4 つである。このうち本研究が対象とするのは(1)である。これは，「始まる」「起こる」など聞き手の認識が未だ及んでいないであろう事物を談話に導入する働きを持つ動詞(多く自動詞)を指す。以下に例を示す(寺崎 1982: 102-103)。

Empezó	*la*	*resistencia.*
始まる.点過去.3 人称単数	定冠詞.女性.単数	抵抗運動.女性.単数.主格

「抵抗運動が始まった(著者訳)」

この現象に関しては，具体的にどの動詞が「提示型動詞」に含まれるのか，主語後置の生産性はどの程度なのかなど，曖昧な点があり分析・探究の余地がある。

2.2. 先行研究の概観

スペイン語は屈折性が強く，動詞の語形から主語の人称・数が判別できる。そのため主語の省略・後置のハードルが低いと言え，語順は「自由選択」「無秩序」という考え方が長く支配的であった(Contreras 1976: 15)。しかし，Hatcher(1956)が「開始(*beginning*)」，「発生(*occurrence*)」などの意味素性を設定し，主語後置が起こりやすいとされる動詞を類型化したことを契機に関心が強まった。1970 年代には Contreras(1976)のように「主題(*theme*)—評言(*rheme*)」の枠組みなどの情報構造概念の援用もなされるようになった。

上述のような機能主義的枠組みに依拠した統計的研究もある。語順を類型化して計量するという意味で先駆的なのは出口(1984)であった。2010 年代には，線形重回帰分析を用いた研究も発表された。Brunetti & Bott(2011)や Rivas(2013)が該当し，いずれも「情報の新旧」や「項の意味役割」などの機能主義的変数を含めて解析がなされた。しかし，いずれも事例数が小さく，また多重共線性の確認(分散拡大係数の調査など)がなされていないなど，統計学的妥当性の検証が徹底されていない面が見られる。

3. リサーチデザイン

3.1. 研究目的と研究設問

先行研究は，機能主義的な説明変数を採用しているために，その手間から事例数が小さくなり，かつ変数コーディングに恣意性が入りかねないという問題を抱えている。そこで本研究では，機能主義的な説明変数をすべて捨象し，表層から得られる情報のみを利用して機械学習モデリング(ロジスティック回帰)を行うこととした。設定した研究設問は以下である。

(1) スペイン語の主語位置にみられる傾向性を，表層特徴から定量的に説明・予測できるか
(2) スペイン語母語話者が「自然に響く」と感じる語順の背後の要因は何か
(3) 表層特徴と機能主義的概念(情報構造)は，どのように対応しているか

事例数を増やすことと，先行研究(線形重回帰)よりも柔軟なモデリング(ロジスティック回帰)を行

うことで，より効果的に(1)を解決する。これは，より一般的な問いである(2)や(3)への足掛かりとなる。「対象構文(3.3.2 項)」を今後広げていくことで，モデルを敷衍し，(2)のようなスペイン語学的な知や，(3)のような一般言語学に還元できる知を得ることを視野に入れている。

3.2. データ

Cristian Cardellino 氏が公開(Cardellino 2019, Accessed: July 9, 2021)している，20 億語コーパス(述べ語数:20 億 2495 万 9560 語，異なり語数:915 万 7394 語)を用いた。これは The Open Parallel Corpus Project という，話し言葉・書き言葉双方が収録されたコーパスデータに，アルゼンチンの法律文や議会議事録などを加えて集成したデータである。

3.3. 手法

3.3.1. データの事前処理

コーパスの元データは，表 1 に示した通り，品詞タグなどの付いた TSV 形式である。

表 1 元データの一例

0	Reglamento	Reglamento	PROPN	WSP
1	de	de	ADP	WSP
2	Ejecución	Ejecución	PROPN	WSP
3	…	…	…	…

このデータから，Python3 の前処理プログラムで表層語形を抽出して 1 行 1 文形式のテキストに直し，対象構文だけを取り出した。なお，5 語以上の文長の重複事例は，慣用表現にしては長すぎると判断して削除した。前述の「対象構文」とは，「主語」「自動詞」「付加語句」からなる単文を意味している。そのため，対格・与格・再帰代名詞・従属節・関係節・受動態・繋辞を含む文と感嘆文・疑問文は除外した。また，明示主語を持たない事例や，主語が 1，2 人称の事例は除外した。1，2 人称での明示主語は殆ど代名詞のみになってしまい，極端な分布になるためである。こうした過程を経て，最終的に 28,396 個の事例を収集した。なお 3.3.1 項，3.3.2 項の各処理には，Explosion 社が Python3 向けに提供する spaCy ライブラリ(3.0.6)の es_core_news_lg モデル(3.0.0)を使用した。このモデルは，構文解析などの正解率が軒並み 9 割前後(spaCy, Accessed: August 20, 2021)と，信頼性が高い。

3.3.2. 分析の手順

前処理済みのデータセットから，目的変数と 25 種類の説明変数を自動抽出し，データを 3:1(訓練データ:評価データ)に分割した。目的変数は，主語後置の場合「1」，前置の場合には「0」というダミー変数として扱った。説明変数(表 2)のうち，PrevPresence(当該文の主語が，直前 3 文のうち何文に登場していたか)と PostRate のみが量的変数であり，その他は，該当時に「1」，さもなくば「0」のダミー変数として扱った。なお PostRate は，訓練データからのみ算出することで訓練データから評価データへの情報転移を防いだ。

学習段階では，scikit-learn 0.24.2 を利用して訓練データからロジスティック回帰モデルを構築した。2 つの説明変数間の交互作用項も考慮したので，次元数は 325 になった。正則化は，説明

表 2 説明変数一覧

説明変数	略号
直前 3 文の状況	PrevPresense
主語が複数	Plural
主語が固有名詞	Proper
主語が人称代名詞	Pronoun
主語が指示詞(を含む)	Demons
主語が属格(を含む)	Genitive
主語が量化詞(を含む)	Quantifier
主語に定冠詞あり	DetArt
主語に不定冠詞あり	IndArt
当該文が否定文	Negative
述語動詞が直説法	Indicative
述語動詞が接続法	Subjunctive
述語動詞が現在完了	PresPerf
述語動詞が過去完了	PastPerf
述語動詞が現在進行	PresProg
述語動詞が過去進行	PastProg
述語動詞が点過去	SimplePret
述語動詞が線過去	Imperfect
述語動詞が未来	Future
述語動詞が接続法過去	Past(SBJC)
述語動詞が過去未来	PosPret
述語動詞が現在	Present
文頭に前置詞句あり	InitialPrep
文頭に副詞(句)あり	InitialAdverb
述語動詞の主語後置率	PostRate

変数ベクトルの L2 ノルムに定数項を掛けたものを罰則項とすることで実行した。このとき，定数項の逆数 C の値を 1.0×10^{n}(n= -2, 0, 2, 4 ,6)の要領で変化させ比較した。そのうえで，各モデルにおいて評価データに対して主語位置の予測(分類の閾値は 0.5)を実行した。

4. 結果と考察

まず，構築したモデルの情報を示す。表 3 は，訓練事例・評価事例に対する正解率・適合率・再現率(有効数字 3 桁)である。比較しうる先行研究として，Brunetti & Bott(2011: 28)が構築した決定木のスコアも併記した。表 4 は，最も優秀なモデル(C= 1.0×10^{4} のとき)の各説明変数の偏回帰係数(有効数字 3 桁。「&」は交互作用)である。紙面の都合上，正の値・負の値それぞれ上位 4 個のみを挙げた。

4.1. RQ1:「スペイン語の主語位置にみられる傾向性を，表層特徴から定量的に説明・予測できるか」

表 3 を見る限り，表層特徴を用いて主語後置／前置を予測しうる蓋然性は高いと言える。先行研究(Brunetti & Bott 2011: 28)と比しても，ほとんどの指標で上回っているし，そもそも先行研究では訓練・評価のデータ分割が行われていないので，指標の信頼度は本研究を下回ると考えられる。機能主義的カテゴリを導入せずとも，事例数を増やせば主語位置を正確に予測しうることが明らかになった。

一方で，訓練データと評価データとの間で性能差が小さいことや，正則化を緩くしたときに性能が上がり収束傾向を見せることから，今回のモデルは適合不足だと考えられる。次元を増やしたり，ランダムフォレストなどのフィッティングの強いモデルを適用したりして，さらに分類性能を高めることが可能だろう。

表 3 構築したモデルの評価指標値

先行研究 （事例数 1200）	本研究（訓練事例数 21297，評価事例数 7099）				
	C= 1.0×10^{-2} （正則化最高）	C=1.0	C=1.0×10^{2}	C=1.0×10^{4}	C=1.0×10^{6} （正則化最低）
＜訓練＞ 正解率:.838 適合率:.736 再現率:.349 ＜評価＞ なし	＜訓練＞ 正解率:.803 適合率:.772 再現率:.638 ＜評価＞ 正解率:.815 適合率:.789 再現率:.641	＜訓練＞ 正解率:.811 適合率:.764 再現率:.681 ＜評価＞ 正解率:.815 適合率:.767 再現率:.676	＜訓練＞ 正解率:.812 適合率:.758 再現率:.696 ＜評価＞ 正解率:.815 適合率:.760 再現率:.685	＜訓練＞ 正解率:.812 適合率:.758 再現率:.696 ＜評価＞ 正解率:.815 適合率:.760 再現率:.686	＜訓練＞ 正解率:.812 適合率:.758 再現率:.696 ＜評価＞ 正解率:.815 適合率:.760 再現率:.686

4.2. RQ2:「スペイン語母語話者が「自然に響く」と感じる語順の背後の要因は何か」

表 4 偏回帰係数の絶対値が大きい説明変数（C=1.0×10^{4}のとき）

説明変数	偏回帰係数（正）
Past(SBJC) & InitialPrep	9.91
Demons & Quantifier	6.88
Demons & Subjunctive	6.56
PastProg & InitialAdverb	5.48

説明変数	偏回帰係数（負）
Pronoun	-8.24
DetArt & PastProg	-7.64
IndArt & PastProg	-6.44
Genitive & PastProg	-6.39

ロジスティック回帰は一般化線形モデルの一種であるから，表 4 のように説明変数ごとの偏回帰係数の符号と絶対値を見ることで，主語位置の分類におけるインパクトを推し量ることができる。表 4 では，正の値（主語後置へのインパクト）での上位 4 つはいずれも交互作用項だった。過去進行（PastProg）や指示詞（Demons）などが主語後置へのインパクトが強い可能性がある。ただし，とりわけ「過去」の表現に関してテンスやアスペクトが複雑に入り組んだパラダイム（表 2 参照）を持つスペイン語において，過去進行形は英語ほど頻繁に使われない。少数の事例から強い影響を受けている可能性も高いので，正則化などを通じて実験を繰り返す必要がある。

負の値（主語前置へのインパクト）では，依然として過去進行の影響が強いが，一方で人称代名詞（Pronoun）や定冠詞（DetArt），属格（Genitive）などが主語前置に作用している可能性が見て取れる。特に定冠詞は，野田（1994: 94-95）などに見られるよう，「情報の新旧」における旧情報に振り分けられる傾向にある。こうした先行研究との比較検討が重要になるだろう。

また，後置率（PostRate）の値は，Hatcher（1956）らの「動詞の意味素性」との比較検討が可能である。例えば Hatcher（1956: 8）は「開始（*beginning*）」という素性を立て，この素性を持つ動詞は主語後置が無標になると指摘した。しかし，本研究においては「始まる」という意味を持つ *empezar* や *comenzar* はいずれも 2 割前後の後置率（それぞれ 2.43×10^{-1} と 1.94×10^{-1}）にとどまっている。この解釈については，今後検討する必要がある。

4.3. RQ3:「表層特徴と機能主義的概念(情報構造)は, どのように対応しているか」

4.2 節で示した通り, 大規模な定量分析を経て, 表層特徴(定冠詞の有無など)から「情報の新旧」などの機能主義的概念が近似できる可能性が明らかになった。また, 本研究では時制などの情報や, 交互作用項も考慮しているので, 例えば「定冠詞 & 現在形」と「旧情報」の関連を考察するなどの, よりきめ細かい分析をすることが可能となる。さらに, Contreras(1976)らの立場に倣い, 後置主語を「評言」・「新情報」と仮定するならば, 各事例文の持つ説明変数ベクトルと主語位置との間でコレスポンデンス分析を行うなどにより,「典型的な主語後置構文」の類型を見出し, 間接的に表層特徴と情報構造の関連を調べることもできるだろう。

5. まとめ

本稿は, スペイン語の自動詞構文における主語後置のメカニズムに, 大規模データからの数理モデル構築と, その機械学習結果の言語学的解釈で迫ろうとした。研究設問の(1)～(3)について, 結果から, (1)では主語位置を表層特徴から予測できる可能性は高いと明らかになった。(2)では, 先行研究との一致・不一致の両方が見られた。(3)では,「定冠詞」などの特徴が「情報の新旧」などと対応するという解釈が可能であった。これは, 野田(1994: 94-95)らのスペイン語学的知識に符合している。他方, 少数の事例から強い影響を受けていると解釈できる箇所もあった。正則化手法など, アルゴリズムの修正が必要となる。また, (3)を通言語的に敷衍できれば, 一般言語学的に有益たりえる。

引用文献

Brunetti, L., &. Bott, S. (2011). *Subject inversion in Romance: a corpus-based study*; Handout distributed at: Quantitative Investigations in Theoretical Linguistics QITL-4. <available at https://edoc.hu-berlin.de/bitstream/handle/18452/2021/brunetti.pdf?sequence=1>

Contreras, H. (1976). *A theory of word order with special reference to Spanish.* North-Holland Pub. Co.

Cardellino, C. (2019). *Spanish Billion Words Corpus and Embeddings* (Published: August 2019), <https://cs.famaf.unc.edu.ar/~ccardellino/SBWCE/sbwce.tagged.tar.bz2>(Accessed: July 9, 2021)

出口厚実 (1984)「スペイン語における主語・動詞・目的語の語順に関する量的考察」『Estudios Hispánicos』(大阪外国語大学), *10*, 1-17.

Hatcher, A. G. (1956). Theme and Underlying Question: Two Studies of Spanish Word Order. *Word 12,* supplement *3*, 1-52.

野田尚史 (1994)「日本語とスペイン語の無題文」『日本語とスペイン語』(国立国語研究所), *1*, 83-103.

Rivas, J. (2013). Variable Subject Position in Main and Subordinate Clauses in Spanish: A Usage-Based Approach. *Moenia, 19*, 97-113.

spaCy. *spaCy Models Documentation: Spanish* < https://spacy.io/models/es> (Accessed: August 20, 2021)

寺崎英樹 (1982)「スペイン語における主語の後置」『小樽商科大学人文研究』, *63*, 95-111.

『英語コーパス学会大会予稿集 2021』(ISSN 2436-6447)
Proceedings of the JAECS Conference 2021

刊行日 2021 年 10 月 2 日
発行所 英語コーパス学会
事務局 〒819-0395 福岡市西区元岡 744 九州大学大学院言語文化研究院 内田諭研究室気付
jaecs.hq@gmail.com