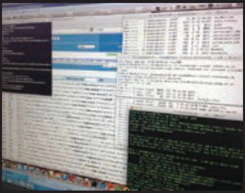
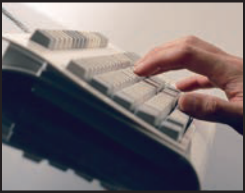
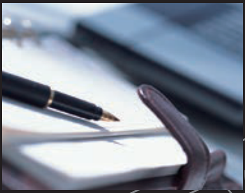
JAECS
Japan Association for English Corpus Studies

2023
英語コーパス学会

ISSN 1340-301 X



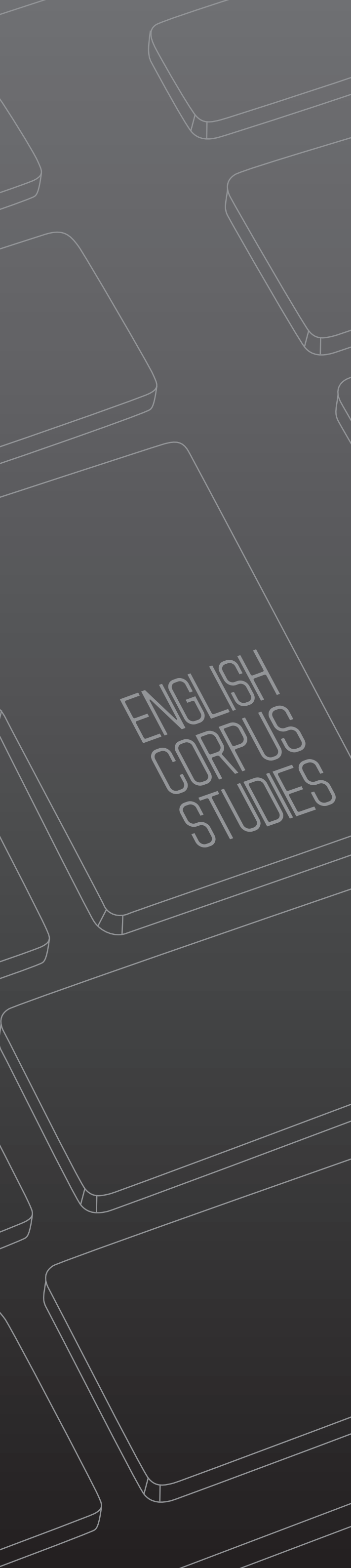




英語コーパス研究
第30号

ENGLISH
CORPUS
STUDIES

English Corpus
Studies 30



2023年度
英語コーパス学会役員

ISSN 1340-301 X

英語コーパス研究

第 30 号

英語コーパス学会

2023

目　次

記念号巻頭言

論文

研究ノート

資料

巻頭言

『英語コーパス研究』第 30 号（英語コーパス学会創立 30 周年記念号）刊行に寄せて

田畑　智司

本学会創立の翌年，平成 6 年（1994 年）に創刊された『英語コーパス研究』がこの度，第 30 号を迎えることは，まことに喜びに堪えません。本号を含め，これまでに掲載された原著論文は 98 篇，研究ノート 36 篇，誌上シンポジウム論文 67 篇，特別寄稿・講演 13 篇，コーパス・ソフトウェア紹介 12 篇，書評・図書紹介 10 篇，海外リポート 5 篇，教育実践報告 1 篇，合わせて 242 篇にのぼります。また，本誌は巻末に当該年度の大会プログラムと資料を記載しており，当学会の歩みを記す貴重な年譜ともなっています。

この 30 年間に掲載された論文や記事を振り返ると，最初の 10 年は語法・文法や通時的変異，史的言語研究，多変量解析によるテクスト類型論などを扱った論文に加えて，コーパスおよびソフトウェア紹介や，書評・図書紹介，海外リポートが多く掲載されているなど，黎明期にあった当学会の姿が如実に反映されているように思えます。次の約 10 年は，誌上シンポジウム論文が積極的に公表されるようになる一方，掲載論文で扱われるトピックの範囲が大きく広がりました。特に，パラレルコーパスや学習者コーパス，母語話者・非母語話者による言語産出を収録した対照コーパスなどの構築や活用に基づき，学習者やアカデミックライティングの言語特徴を明らかにする英語教育学関連の論文が増加したほか，コーパスとしての Web，認知言語学におけるコーパス利活用，特定の言語使用域や集団の特徴抽出の方法論などが盛んに議論されるようになり，本学会の発展の様子がうかがい知れます。

さらに，最近の 10 年余りでは，データ駆動型学習，ESP や EAP，アカデミッククライティングのムーブ分析，言語処理を応用した分析や作文を支援する新たなツールおよびシステムの開発，そして理論言語学・認知言語学とコーパス研究との融合などから見て取れるように，コーパス研究はさらに発展展開を続けてきました。なかでも，機械学習に基づく分類器や特徴（量）抽出，そして，トピックモデリングや word2vec など，急速に進歩する自然言語処理の知見がコーパス研究に導入されるようになってきたのも新たなトレンドだと言えるで

しょう。

この間，コーパス研究の外側では，自然言語処理が長足の技術的進歩を遂げています。とりわけ，Bidirectional Encoder Representations from Transformers (BERT) や Generative Pre-trained Transformer (GPT)-3/GPT-4 に代表される，transformer を基盤とした深層学習による大規模言語モデルは，正確な数値は公開されていないものの，極めて巨大な文章データをコーパスとして学習させています。その結果，爆発的な勢いで汎化性能を高め，生成 AI としてまさに時代の寵児となる一方，研究倫理や規範，著作権，個人情報に対する脅威とも目されるほどに高度な文章生成能力を備えていることは，もはや人口に膾炙するところとなりました。

生成 AI は文章校正や，参考文献リストの整理・確認，プログラミングコード生成など，すでに学界でも利用が始まっています。他方，研究倫理の面からその利用ガイドラインが議論され始め，本誌の投稿規定にも生成 AI の利用についての条項が追加されることが決まっています。私たち英語コーパス研究にたずさわるものにとって，これからの学術研究は好むと好まざるとに関わらず，AI と共生・共存していくことになるでしょう。これまで，私たちはコンピューター，プログラミング，統計学，言語処理などをスキルとして味方にする術を模索してきました。そして，コーパスを通して初めて可視化される言語使用の実際を観察し，記録することにより，理論や仮説の発見・検証，さらには新たな問いの創成など，知の蓄積と継承に取り組んできました。そのことは 30 年にわたって本誌に掲載されてきた論考に克明に刻まれています。生成 AI をうまく利用することにより，コンピューターにさせる仕事と，人間がすべき仕事―着想，分析，議論，そして洞察―とを適切に切り分け，私たちが後者に専念し，その質をより一層高めるためにはどうすればよいのか。本誌『英語コーパス研究』が，まさにその答えに至る道標を照らす論考の宝庫（アーカイヴ）となることを大いに願う次第です。

（田畑　智司　大阪大学）

「論文」

‘To spread the Word by which himself had thriven’: Analysis of Alfred Tennyson’s Use of Language Based on the LDA Topic Model

Iku FUJITA

Abstract

This study aims to provide pieces of newly discovered dimensions on the poetic style of the Victorian poet Alfred Tennyson, using a stylometric approach, the latent Dirichlet allocation (henceforth, LDA; Blei et al., 2003) topic model. Many studies have examined Tennyson’s poetry by focusing on similarities or differences between Tennyson’s and other poets’ styles, syntax or lexicons. However, most take a qualitative approach to specific poems based on close reading of texts. This study attempts to uncover new aspects of Tennyson’s style through a balanced combination of close reading and quantitative analysis. LDA is a method dedicated to analysing big data, making it possible to read text data from a ‘distance’. Moreover, ‘distant reading’ enables us ‘to focus on units that are much smaller or much larger than the text: devices, themes and tropes’ (Moretti, 2013: 48–49). Using probabilistic calculations, LDA identifies semantic connections hidden behind words in the target corpus. Although several studies have employed the topic model to investigate prose texts, very few have applied the technique to analyse poetry. However, Fujita (2022) showed that LDA is effective in examining Tennyson’s poetry. This study examines the lexicon of Tennyson’s 603 works, focusing particularly on the nouns he uses. Results of this analysis show that some topics (semantically classified word groups), such as ‘immortality’, one of Tennyson’s major themes, are the most prevalent in his body of work.

1. Introduction

Alfred Lord Tennyson (1809–1892) is one of Britain’s representative poets of the

Victorian period. He became a poet laureate in 1850, after the death and the resignation of the previous laureate, William Wordsworth (1770–1850). Tennyson wrote more than 600 poems, including unpublished ones, of which more than 80 percent are lyrical poems. However, his major works are narrative and epic poems, which tend to be longer than lyrical poetry. One of his famous lyrical poems, *In Memoriam A.H.H.* (1850; henceforth *In Memoriam*), consists of more than 2,800 lines. It can be said, then, that one feature of Tennyson's body of work is his long poetry.

Numerous past studies exist on Tennyson's life and his poetry, style, prosody and lexicon (e.g., Dixon, 1896; Tennyson, 1897; Nakamura, 1967; Ricks, 1969, 1987; Nishimae, 1979, 2000; Bloom, 1985; Hair, 1991; Kabata, 2001, 2007; Noguchi, 2011; Holmes, 2012; Thomas, 2019), and much of these studies adopt a qualitative approach to specific poems. Plamondon (2005) is the sole study that takes a quantitative approach and investigates the difference in some sounds' frequencies between Tennyson's and Robert Browning's poetry. However, other than Plamondon's research, few Tennysonian studies employ quantitative approaches amongst extant studies. The qualitative studies above point out or investigate topics such as 'Auditory Value in *Enoch Arden*' (Horiuchi, 1992), 'Tennyson and Death' (Bruce, 1917), and 'Tennyson and Spiritualism' (Elliott, 1979). Horiuchi (1992) offers detailed observations of sounds and prosody in *Enoch Arden* (1862), as reflected in the title of his paper. Bruce (1917) and Elliott (1979) discuss the themes of death and spiritualism respectively. Bruce (1917) focuses on poems such as 'Lover's Tale' (1833), 'Morte d'Arthur' (1842), 'St. Agnes' Eve' (1836) and 'Gareth and Lynette' (1872), while Elliott (1979) spotlights comments for Tennyson's works from Tennyson himself and his close people, such as Frederick, to explore his works like 'The Ancient Sage' (1885), 'The Silent Voices' (1892) and *In Memoriam*. Tennyson's 'Tithonus' (1862), *In Memoriam* and 'Crossing the Bar' (1892) often attract attention from literary critics in examinations of the subject of immortality in his poetry (e.g., Shaw, 1976; Elliot, 1979; Perrine, 1966). Ricks (1987) is a representative study that comprehensively annotated Tennyson's works with abundant materials. Various critical studies refer to Ricks (1987), or to the earlier edition, Ricks (1969), and Shaw (1973) and Hair (1991) are no exception. Shaw (1973) discusses the similarity between Tennyson's syntax and style and that of other poets and playwrights, for example, Homer, William Shakespeare, John Keats, Wordsworth and others, by identifying and counting the number of allusions to other authors, as identified in Ricks (1969). Hair (1991) illustrates qualitative features by comparing words, concepts or motifs with those of other poets. A recent critical study by Thomas (2019) employs close reading to point out the *echoes* in Tennyson’s works of his predecessor as poet laureate, Wordsworth.

What these earlier studies have in common is not only that they undertake detailed and in-depth close readings, but also that they focus on specific words, phrases or works. However, there is rarely critical research that presents an overall picture by observing an entire work(s). In other words, although there are much *other elements* we might wish to see outside of a specific/representative work, we are missing them by focusing on only a limited part of Tennyson’s works.

Moretti (2013: 48) explains, ‘the trouble with close reading […] is that it necessarily depends on an extremely small canon’ in the literature research. He continues, saying close reading leaves ‘great unread’ in works in question since it focuses on a limited part of work(s). Moretti adds, ‘if you want to look beyond the canon’ (2013: 48), ‘distant reading’ will work effectively for this. His ‘great unread’ corresponds to the abovementioned *other elements*. Though Moretti does not specify what the ‘canon’ encompasses, his remarks can be applied to research on individual poets/authors. Nonetheless, the intent of Moretti (2013) and the present paper is not to criticise studies based on close reading or the approach itself. *Other elements*/‘great unread’ can cover a wide range, but this study focuses on two divergences: the possibility of previous studies might have missed and not discussed poems, which contain the same themes as other Tennysonian canons; the prospects of unmentioned topics by earlier scholars, though written within his poems. More precisely, as mentioned in previous studies, even though immortality has been a general theme, the poems in question were supposedly limited to the works mentioned above, such as *In Memoriam*, ‘Crossing the Bar’ and ‘Tithonus’. Likewise, some studies indicate relationship between seas and emotions of characters as well as what seas connote in his particular poems, for instance ‘Mariana in the South’ (1832), ‘Œnone’ (1832), ‘The Mermaid’ (1830) and ‘Morte d’Arthur’ (Fulweiler, 1965; Keirstead, 2019). Meanwhile, Okazawa (1969) points out Tennyson’s preference of waters, namely seas, rivers and lakes; rivers have not been fully recognised by Tennysonian scholars compared to seas. *The Princess* (1847) and some other works have also been given earlier attention in the way Tennyson depicts female characters; however, male protagonists have not been noticed as much as their female counterparts have. Because few studies exist that employ distant reading, including statistical methods and stylometric approaches, this paper aims to reveal the 'great unread' in Tennyson's poems through a quantitative method, the latent Dirichlet allocation (LDA) topic model (Blei et al., 2003).

The basic idea behind LDA is that 'documents are represented as random mixtures over latent topics, where each topic is characterised by a distribution over words' (Blei et al., 2003: 996). LDA classifies words into groups, which are called topics. As Tabata (2018: 52) indicates, 'LDA is a machine learning method for uncovering hidden semantic structures in a corpus of texts'. The main advantage of using LDA is that this approach allows to find semantic connections between words in texts, which other quantitative methods like keyword analysis, find it difficult to identify. LDA is especially dedicated to analysing big data, and it allows us to read text data from a 'distance'. Using probabilistic calculations, LDA detects the semantic connections hidden behind words in the target corpus. Since LDA is more efficient algorithm on big data (Iwata, 2015), several studies have employed the topic model to examine prose texts (Onodera et al., 2016; Kuroda, 2017; Tabata, 2017, 2018, 2020; Kiyama, 2018; Matsukawa et al., 2018; Huang, 2020a, b), yet relatively few have applied the technique to analyse poetry (Rhody, 2012; Navarro-Colorado, 2018; Henrichs, 2019; Okabe, 2019). Although this might be perceived as a challenge for applying LDA on poetry, which is basically short-length compared to prose texts, Fujita (2022) shows that LDA is effective in considering Tennyson's poetic works. Referring to Fujita (2022), this study investigates the lexicon of Tennyson's 603 works, focusing in particular on the nouns he uses. The emerging LDA results show that some topics (semantically classified word groups) appear in multiple poems. For example, a topic represents the factors of 'immortality', one of the key themes in Tennyson. Further, this method reveals multiple associations with the word 'man' throughout his works. New 'unread' dimensions can be uncovered by a balanced combination of close reading and emerging results of quantitative analysis.

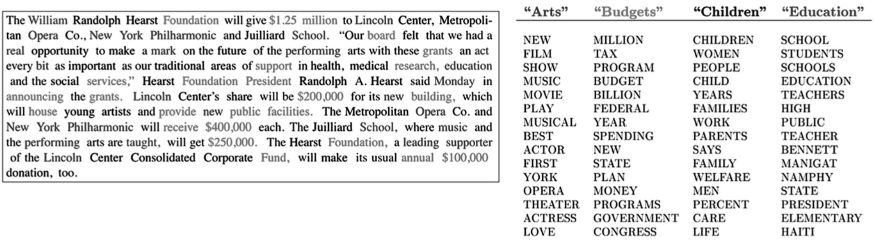
2. Methodology and data

2.1 LDA topic model

The LDA topic model calculates probabilities and classifies words into topics based on the hypothesis that documents should be classified into several groups based on word co-occurrence (in the same document) trends. Here 'document(s)' is defined as consecutive segments of text data for analysis but that are not always relevant to a poetry/prose work (see Section 2.2 for more detailed explanation on slicing texts into consecutive segments of an equal size).

Figure 1 shows that, in running LDA, the passage at the left of the figure is analysed and the words are classified into groups called 'topics', and four topics as well as their constituent keywords are presented at the right of the figure.

LDA adopts a *bag of words* model, which means word order is ignored, unlike collocation and *n*-gram, which primarily take into account the order of words in a row. As mentioned previously in the first section, LDA uncovers a semantic connection between words; however, LDA does not draw on a semantic classification dictionary as Wmatrix (Rayson, 2009) uses. Therefore, it is the analyst who has to interpret outputs of LDA by combining their knowledge about the text data and assigns labels for each topic, for example, 'Arts', 'Budgets', 'Children' and 'Education' in Figure 1.



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Figure 1. Example article from the AP corpus (adapted from Blei et al., 2003: 1009, modified)

2.2 Data and settings for analysis

This study analyses 603 poetical works written by Tennyson. Each work was collected as it appeared in a first edition or, if not published, as first written. Of the 603 poems, 423 were compiled as part of the Delphi Poets Series 'Alfred, Lord Tennyson' (2013), and the other 180 were compiled by Ricks (1987). Minute but detailed emendation is given manually on digital documents converted by optical character reader/recognition. Table 1 shows statistical data of the Tennyson corpus.

Prose texts have been the subject of previous LDA studies. The number of studies

Table 1. Statistical data of Tennyson corpus

The number of poems	603
Total tokens	358,190
The shortest poem in number of words	12
The longest poem in number of words	26,383
Mean tokens per poem	594.01
Standard deviation	1,823.15

on poetry texts is extremely few compared with studies dealing with prose works. Poetry texts typically have fewer tokens per work than prose, which in part accounts for why there are not many studies of poetry using the LDA method. Further, LDA requires a significant number of documents for its application. Although the data size of the Tennyson corpus is not as large as those of some other prose text studies, the data for the topic model are considered sufficient, not only because the Tennyson corpus contains 603 poems but also because the total token of the corpus is over 350 thousand by reason of Tennyson's several lengthy poems, such as *The Princess*, *In Memoriam* and *Maud* (1855).

In LDA, which 'treats each document as a *bag of words*', document size deviation is equivalent to bag size variation, and 'the bigger the bag, the more words that will tend to be found together in the same bag' (Jockers, 2014: 137). With regard to the study of prose text, Jockers (2014: 137) explains that LDA captures 'some themes that run throughout and others that appear at specific points and then disappear in novels' and that 'it is useful to divide novels (and other large documents) into *chunks* or *segments* and then run the model'. Prose texts from earlier research that employed LDA, have been sliced into segments of equal size, such as 1,000 (Kuroda, 2017; Tabata, 2020), 2,000 or more words (Huang, 2020b). Splitting the poems in the Tennyson corpus into equally sized segments is preferable considering the lengthy poems in the corpus. Further, the large disparity in *bag* size, namely segment size, instantly affects the results since LDA uses raw word frequency to identify the word co-occurrence. In Table 1, the standard deviation value shows that Tennyson's poems are of varying lengths. The high degree of variation in total word tokens in texts represents a potential methodological issue: a larger number of words may cause over-representing the topic since longer texts are more likely to contain a larger number of words belonging to a particular topic, compared with texts with an equal proportion of words for the same topic. Therefore, it is necessary to ensure texts to be fed into the topic modelling of a reasonably comparable size, if not exactly the same. In regards to these concerns, Fujita (2022) suggests the relevant size of the segment for LDA practice in Tennyson’s poems. This paper refers to Fujita (2022) and adopts a segment size of 594 words by the mean token value of the Tennyson corpus (Table 1). A segment is not exactly the same to one poem, but they correspond. For the poems that contain more than 594 words, for example, *The Princess*, of which total tokens are 26,526, is split into 45 consecutive segments, sequentially counting the number of tokens from the beginning (first token) to the end of the poem. When each text was divided into equal-sized consecutive segments, the two final parts were joined unless the final chunk was 12 word-long, the same length with the shortest poem in the Tennyson corpus (Table 1). Additionally, the poems with total tokens of fewer than 594 words will not be divided into duplicate segments but will be treated as a segment per work. Thus, the largest segment size is 606-word length, and the smallest segment size consists of 12 words.

All words in texts were assigned part-of-speech tags using CasualConc 2.0.8 (Imao, 2022). This paper used a tagset called CLAWS5, as given in the British National Corpus. In order to consider prominent concepts, themes and subjects (all of these can be incorporated into topics discovered), the present study confines its scope to nouns used in the texts in the hope that nouns are more likely to encapsulate the ideational content of a text than other parts-of-speech. After the works were separated into 594-word consecutive segments, other words besides nouns were deleted using the part-of-speech tags; thus, LDA only analyses the target nouns.

LDA was applied to the segments using the MAchine Learning for LanguagE Toolkit (McCallum, 2002). The number of topics was set at 20 based on consideration of emerging results of prior experimental trials, with the number of topics ranging from 10 to 200.

3. Results of LDA

This section discusses the results of LDA for the Tennyson corpus. Table 2 shows one of the output results of topic modelling, indicating the labels and alpha (α) value attached to each of the 20 topics, and the top 20 most salient keywords for each topic. The keywords are ranked in descending order of their weights from the top to the bottom of the table. The α value represents the universality of each topic. The higher the value, the more prominently the topic appears in multiple works; the lower the value, the topic appears in a small number of works, sometimes only in a single work. In Table 2, the most prominent topic with the highest α value is Topic 17 (α = 0.88347). In contrast to Topic 17, Topic 2's α value (α = 0.03632) is the lowest among the 20 topics. Thus, it appears that the top 20 keywords of Topic 17 are the most significant elements, which appear in multiple works within the target corpus.

The labels in Table 2, are not the results that are automatically suggested by the LDA analysis. The author of this paper considered each label in terms of most relevant through a meticulous reading of the poems and the results. Table 2 shows that some topics are included with similar labels, but the keywords of these topics are somewhat different. Topics 17 and 7, the most significant topics, Topics 0 and 1 and Topics 6 and 18 have partial aspects in common. Topics 17 and 7 are both labelled as 'Life,' but Topic 17 is more about 'materials' in life, while Topic 7 relates to 'emotion'. Topics 0 and 1 are twin topics referring to the 'environment' depicted in the poems, but what separates the two topics is whether the 'environment' is near *sea* or *river*. While previous studies mainly focus on *sea* among waters as mentioned earlier, another watery element in nature, *river*, as well as *sea*, was detected from the LDA results. Some topics, Topics 11, 13 and 19, are about women, which, despite having α values that are not high, are associated with female protagonists and help support earlier studies that have highlighted women as constituting a dominant theme in Tennysonian poems.

Whereas Topics 11, 13 and 19 were found for females, Topics 6 and 18 both have males in their labels. Topic 6 refers to a *man* and his association with family members, while Topic 18 reflects relations between *man* and the society to which he belongs. The following section discusses Topics 17, 7, 0, 1, 6 and 18 as three pairs that correspond to each other. With the six topics, the next section further investigates whether other poems that have not been included in the discussion of previous studies have nonetheless been found to contain identical or similar themes as in canonical works, as well as other latent topics that have not gained much scholarly attention.

Table 2. A result of running LDA on the Tennyson corpus; 20 topics with their labels, alpha values and keywords

LABELS	Environment (sea)	Environment (river)	Lives in poems	Religious elements	Nature & beauty	Marriage	Man in family	Life (emotion)	War	Christianity
TOPICS	0	1	2	3	4	5	6	7	8	9
ALPHA VALUES	0.15124	0.07165	0.03632	0.05559	0.23016	0.05058	0.55437	0.77814	0.18978	0.09134
KEY WORDS (1–20)	sea	river	poet	gods	year	bride	man	love	war	fool
	isle	lake	goose	death	rose	moor	child	life	land	flesh
	sail	valley	reviewers	cannon	garden	woman	mother	heart	name	god
	boat	mountain	laurel	gleam	song	ruin	heart	death	battle	sin
	seas	pine	peacock	right	summer	gold	father	soul	glory	fire
	shore	palm	weather	children	flowers	fleet	hand	spirit	throne	mark
	ship	boughs	moment	league	leaves	hall	face	mind	men	priest
	water	brook	world	mountain	flower	cave	wife	eyes	hearts	ringlet
	sand	poplar	volume	roof	leaf	bridegroom	day	thought	voice	church
	melody	fern	stores	anger	birds	ward	life	hope	fame	bread
	ocean	dews	rhyme	valley	air	tide	men	light	freedom	wine
	rocks	fruit	drop	banner	roses	motion	house	tears	trumpet	beauty
	cliff	leaf	hammer	left	love	signs	head	pain	blood	hell
	world	water	chorus	master	morning	cap	boy	memory	kings	friend
	foam	bowers	poverty	brigade	spring	text	children	grief	sons	swine
	blast	grass	sow	charge	winter	fool	son	words	strength	brute
	hills	middle	throng	darkness	tree	beast	word	hour	power	soul
	bay	arches	eggs	people	wood	trifle	woman	things	shame	cross
	moon	summer	laughter	earthquake	woodland	wife	hands	brain	people	saints
	vessel	knees	pint	foe	lawn	water	night	blood	friends	shame

LABELS	Abstracts related to life	Lady & prince	Abstracts on human	Ladies	Object of love	Strength of nature	Nobles	Life (materials)	Man in society	Mother & queen
TOPICS	10	11	12	13	14	15	16	17	18	19
ALPHA VALUES	0.09013	0.0799	0.07364	0.05834	0.21145	0.19772	0.20564	0.88347	0.50529	0.04247
KEY WORDS (1–20)	time	lady	fame	princess	eyes	eye	king	night	man	mother
	sense	prince	charm	woman	love	soul	queen	light	men	queen
	prime	horse	music	prince	heart	earth	knight	day	world	farm
	throne	arms	woman	lady	beauty	woe	knights	sun	time	green
	spirit	knight	dream	women	tears	pride	hall	heaven	things	grave
	blue	wood	note	storm	cheek	air	eyes	voice	truth	dear
	angel	damsel	faith	form	lips	heaven	man	wind	earth	gate
	waters	knave	friend	riflemen	smile	sky	men	eyes	power	brook
	place	rest	deeds	college	morn	flame	maid	earth	peace	rave
	things	hall	birdie	girls	hair	ray	name	fire	days	cow
	depth	squire	others	boys	feet	hills	shield	star	people	maiden
	river	bridge	wits	honour	neck	gale	table	moon	years	new-year
	airs	court	devil	highness	rose	waves	quest	head	race	fellow
	eddies	walls	mood	men	arm	course	sword	time	nature	ill
	glooms	armour	lute	head	kiss	skies	realm	sound	age	cause
	eternity	charger	belt	ladies	brow	realms	court	shadow	life	dale
	raiment	shield	books	south	summer	view	field	stars	times	chestnut
	sides	kitchen-knave	boon	echo	sighs	light	face	death	hands	alleys
	impulse	fight	counsel	court	brows	might	hands	way	faith	daisies
	immortality	helmet	baby	lisette	kisses	glance	word	gold	land	bailiff

4. Discussion

This section discusses whether the LDA results detected any other poems that previous studies have not discussed in their analysis of specific themes as well as other latent topics that scholars have yet to focus on. Findings for the paired Topics 17 and 7 (section 4.1) show that immortality, one of Tennyson's themes, is depicted not just in his canonical works. In fact, Topics 0 and 1 (section 4.2) highlight *sea* as a significant motif in Tennyson's poems that have not been mentioned in previous research along with *river* as a prominent subject that scholars have rarely discussed. Topics 6 and 18 (section 4.3) illustrate how male protagonists are portrayed, whereas females have been more commonly discussed.

4.1 Topics 7 and 17: Emotions and materials related to life

Based on the analysis of the outputs, it was found that Topics 7 and 17 are the most significant topics in the corpus. The highest α value (α = 0.88347) was associated with Topic 17, while Topic 7 had the second highest α value (α = 0.77814) among the 20 topics, slightly less than that of Topic 17. In terms of the top 20 keywords, Topic 17 mainly comprises physical nouns and nouns pertaining to tangible objects perceived through the senses, such as those related to the eyes, ears or skin. Moreover, these nouns represent entities closely associated with human lives (see Figure 3). On the other hand, the keywords of Topic 7 are relatively close to human lives but are mainly abstract nouns linked to people's emotions or thoughts (Figure 2). *Death* is the keyword of both Topics 7 and 17. Words connected to death, such as *grief*, *spirit*, *soul* and *memory*, are keywords of Topic 7, *heaven* is a keyword of Topic 17. Some of these

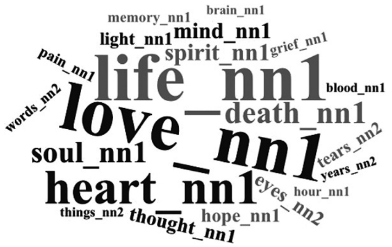


Figure 2. Word cloud of top 20 keywords (Topic 7)

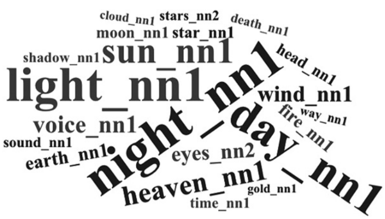


Figure 3. Word cloud of top 20 keywords (Topic 17)

words may not be immediately linked to death, like *soul* or *memory*, but the connection becomes clear when considering notable poetic works related to this topic.

Figures 4 and 5 show the top 50 segments of density of Topics 7 and 17, respectively. Topic density represents the proportion that a topic accounts for in a segment. Segments with higher values have greater topic concentration. In Figures 4, 5 and the forthcoming density plot figures, the top 50 segments, in which the topic densely appears the most, are located horizontally, and the vertical axis shows the topic proportion. Poems containing over 594–606 words were sliced into multiple segments before LDA was implemented; therefore, if a topic significantly appears in multiple segments of one poem, the topic’s proportions are located plurally on a vertical line of the poem. Additionally, when more than three segments of a poem are within the top 50, a box plot appears, and the median value is indicated by a black line inside the box. In these poems, to which the top 50 segments of Topics 7 and 17 belong to, the poet frequently writes about the death of people who are loved and are close[1]. One of Tennyson’s most representative works, *In Memoriam*, is not an exception and is often cited in discussions surrounding immortality in the context of death depicted in a work not only because Tennyson starts with the word ‘immortal’ right at the very beginning, as quoted below, but also because the poem expresses his lamentation over his best friend’s death.

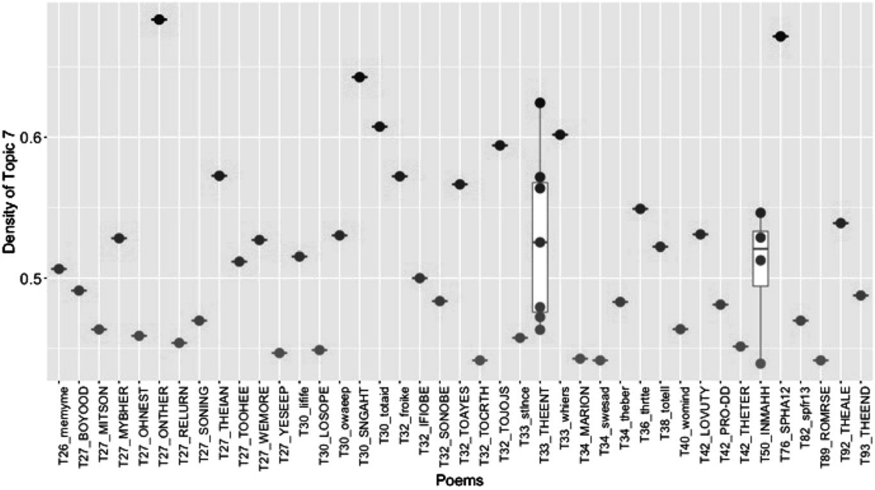


Figure 4. Density bar plot of top 50 segments of Topic 7[1]

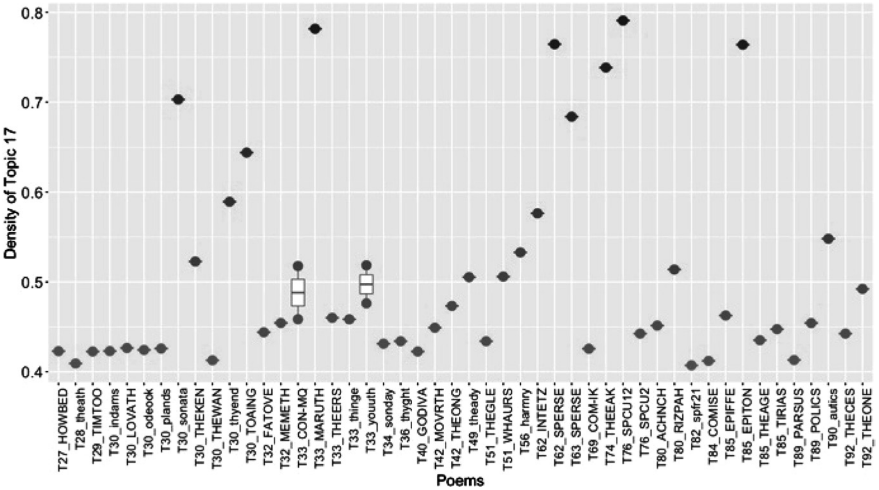


Figure 5. Density bar plot of top 50 segments of Topic 17[1]

Strong Son of God, **immortal** Love,
    Whom we, that have not seen thy face,
    By faith, and faith alone, embrace,
Believing where we cannot prove;

(*In Memoriam A.H.H*, 1850: ll. 1–4; bold added)

The meaning of immortality that is relevant to this poem is '[t]he condition of being celebrated through all time; enduring fame or remembrance' (*Oxford English Dictionary*; *OED*, online, s.v. immortality, *n*. 2.); thus, it does not necessarily mean 'the endless life or existence' (*OED*, s.v. immortality, *n*. 1.). In *In Memoriam* and other significant poems under Topic 7 mention someone who has already passed away; therefore, the focus is not on a person who wishes to obtain eternal life or the depiction of everlasting life, unlike in Tennyson's poem, 'Tithonus', which is often cited in discussions of immortality. Immortality in 'Tithonus,' or in its early original version 'Tithon' (1833), named after a figure from the Greek mythology, is more likely to be seen as expressing the desire for an endless existence. Further, it is not necessarily limited in its meaning of 'enduring fame or remembrance' after someone's death. In fact, the majority of works in Topic 7 explore the death or loss of loved ones, but that is a cue to thinking, remembering and recollecting people who are no longer with the other characters in the poems to some extent. These sensitivities seem to last forever for the people left behind, especially in the immediate moment of grief and despair over a loss. The feelings likely to be eternal are thus considered a form of immortality in the broad sense. While it is possible to identify two different definitions of immortality in Tennyson’s poems, it is suggested, based on the results of LDA, that the poet’s main concern is ‘enduring fame or remembrance’ for someone who has passed away or has simply left the other one’s place.

In the prominent works of Topic 7, immortality is communicated through words like *memory*, *spirit* and *soul*, as well as through a contrast between alive and dead, in addition to using the exact words *immortality*/*immortal*. *In Memoriam*, ‘Crossing the Bar’ and ‘Tithonus/Tithon’ have often been the poetic works that are cited in discussions of immortality in various literary studies (Elliot, 1979; Perrine, 1966; Shaw, 1976; respectively). However, Topic 7 reveals that sense of immortality was a more intriguing concern for Tennyson from his early career until the end of his life, and the theme appears in different poems. Following are two extracts of poems in which Topic 7 significantly appears, ‘To One Whose Hope Reposed on Thee’ (1827) and ‘The Death of the Duke of Clarence and Avondale’ (1892):

> To one whose hope reposed on thee,
> Whose very life was in thine own,
> How deep a wound thy death must be,
> And the wild thought that thou art gone!
> […]
> And **still** I hear the tolling bell,
> For Memory makes each sense her own.
> But **stay**, my soul! thy plaint forbear,
> And be thy murmuring song forgiven!
> Tread but the path of Virtue here,
> And thou shalt meet with her in **heaven!**
>
> (‘To One Whose Hope Reposed on Thee’, 1827: ll. 1–4, 23–28; bold added)

> THE bridal garland falls upon the bier,

The shadow of a crown, that o'er him hung,
Has vanish'd in the shadow cast by Death.
[…]
The face of Death is toward the Sun of Life,
His shadow darkens earth: his truer name
Is 'Onward,' no discordance in the roll
And march of that **Eternal** Harmony
Whereto the worlds beat time, tho' faintly heard
Until the great **Hereafter.** Mourn in hope!
('The Death of the Duke of Clarence and Avondale', 1892: ll. 1–3, 12–17; bold added)

The high α value of Topic 7 suggests its universality in the corpus. Simply put, its elements generally appear in multiple poems. The top 50 segments of Topic 7, as shown in Figure 4, indicate diverse dates of publication, including a considerable number of works from Tennyson's early career. As previously noted, *In Memoriam* and some other poems have been the main works cited in discussions surrounding immortality in Tennyson, but the present results suggest that the poet dealt with immortality in various other poems. Furthermore, it is one of the most familiar themes in his characters' lives along with love and death, which are also always closely linked to human lives.

4.2 Topics 0 and 1: *Sea* and *river*

Topics 0 and 1 have keywords that describe entities related to water, such as *sea*, *river* and *lake* (Figures 6 and 7). The keywords also describe the surroundings of bodies of water in detail. These two topics specifically consist of words related to water in various natural settings, and one of the reasons these keywords to appear in the two topics can be explained by previous studies. Okazawa (1969) points out Tennyson's particular love for and exploration of bodies of waters in nature, such as seas, lakes and rivers. Hair (1991: 42) argues Tennyson was 'a landscape-painter in words, a colourist', pointing out his skill at depicting sceneries through words. Tennyson succeeded in describing and conveying highly detailed scenes to readers, using words, which appear in Topics 0 and 1. Although the LDA results indicated that Tennyson illustrated the things around seas and rivers, previous studies have not frequently argued rivers in

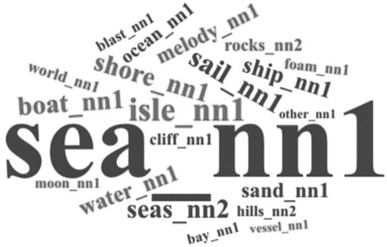


Figure 6. Word cloud of top 20 keywords (Topic 0)

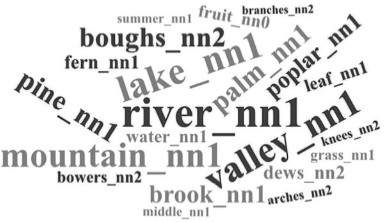


Figure 7. Word cloud of top 20 keywords (Topic 1)

Tennysonian poems. Furthermore, in Topic 0, geographic features are expressed in words such as *shore*, *sand* and *rocks*, which are closely related to the sea, to represent landscapes exquisitely. Watercraft, for instance, *boat*, *ship* and *vessel* also appear, as well as the terms *world* and *seas*, which refer to regions/areas beyond the sea. One possible reason for the appearance of words like some different watercraft, *world* and *seas* in this topic is that Britain is an island country, and to go abroad means to ‘cross the sea’. In contrast, in Topic 1, besides *rivers* and *lakes*, words appear that communicate geographic features, such as *valley* and *mountain*, and also words related to plants, such as *pine*, *poplar*, *leaf* and *grass*. However, words associated with the world or with foreign regions/areas do not appear among the top 20 keywords in Topic 1. This suggests that one of the factors separating Topic 0 and Topic 1 from *rivers* and *seas* is the peculiar nature of Britain, which the sea separates from foreign nations.

The top 50 segments of density of Topics 0 and 1 are shown in Figures 8 and 9, respectively. Based on a close reading of the poems, which the segments appertain to, it is possible to deduce that Tennyson tends to represent feelings of awe, fear, loneliness and despair in the texts given to Topic 0. In scenes where the sea appears, the sound of the waves is almost the only sound that can be heard. The openness of the sea and the absence of any natural or artificial surroundings are also expressed. This nothingness around the person in the scene further amplifies the feelings of loneliness, despair or fear.

Moreover, war is most significant scenes in ‘Jack Tar’ (circa 1859)[2] and ‘The Revenge: A Ballad of the Fleet’ (1878). In these two poems, *sea* is depicted as the place where the war/battle occurs. Combined with images of shipwrecks and war, the sea is associated with the notion of inescapable, unwilling and lonely death. For years,

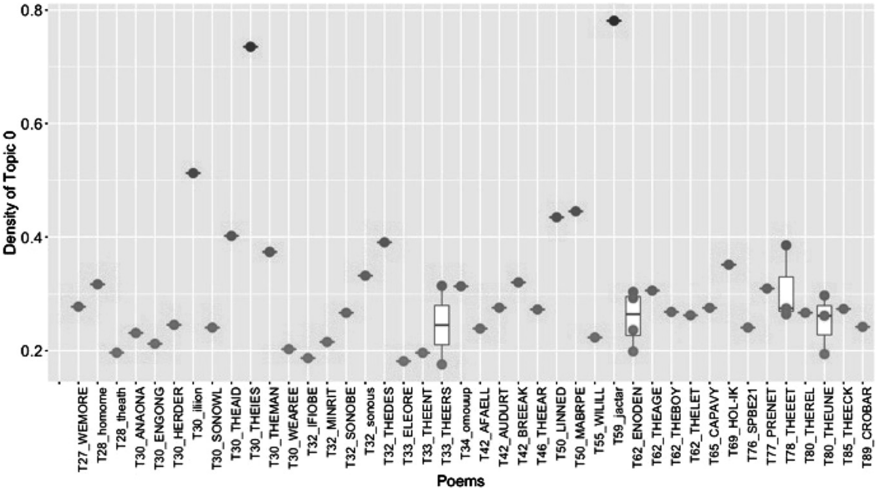


Figure 8. Density bar plot of top 50 segments of Topic 0[1]

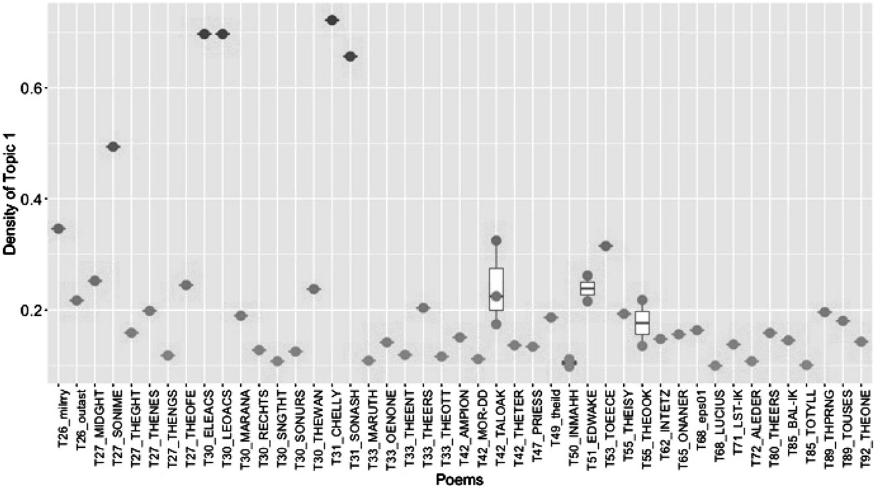


Figure 9. Density bar plot of top 50 segments of Topic 1[1]

scholars have considered *sea* in Tennyson as an attractive theme; however, significant poems in Topic 0, such as ‘Jack Tar’, ‘The Revenge: A Ballad of the Fleet’, ‘Ilion, Ilion’ (1830), ‘Lines [Here Often, When A Child, I Lay Reclined]’ (1850) and ‘Mablethorpe’ (1850), have seldom been cited in the discussion of *sea*. The LDA results suggest a new aspect in some poems outside Tennyson’s canonical or renowned works contain *sea* as their theme.

In contrast, in Topic 1, the anthropomorphic plants and animals talk, sing and surround the people in the poems. Unlike in Topic 0, living creatures and sounds are present in the poems. These works do not convey the impression of fear or despair that is felt in Topic 0’s works. This therefore suggests not only the reflection of the characters’ emotions but also Tennyson’s intention to sketch and evoke these emotions not directly but in an indirect way.

4.3 Topics 6 and 18: *Men*’s various whereabouts

The keywords of both Topics 6 and 18 designate *man*/*men* (Figures 10 and 11). This suggests that the topics are about *men* and some other people and objects that relate to/surround them. The keywords in Topic 6 are words related to *man* and to a man’s family or household, such as *child*, *mother* and *wife*. The top keyword of Topic 18 is also *man*, but the plural form, *men* and *world* closely follow it. While Topic 6’s keywords mainly have a close connection to family and familial life, Topic 18 contains keywords related to man and the world/society outside the home. Among the top 50 segments in which Topics 6 and 18 appear prominently (Figures 12 and 13), nine works in Topic 6 and three works in Topic 18 (within red lines in Figures 12 and 13) are narrative poems (Table 3).

In the narrative poems of Topic 6, the main characters are female, and their relations with families are presented. In contrast, the main characters in ‘Morte d’Arthur’ and ‘Ulysses’ (1842) are powerful, sword-fighting males. The poem titled

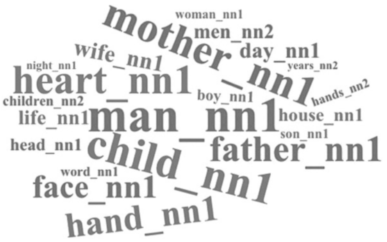


Figure 10. Word cloud of top 20 keywords (Topic 6)

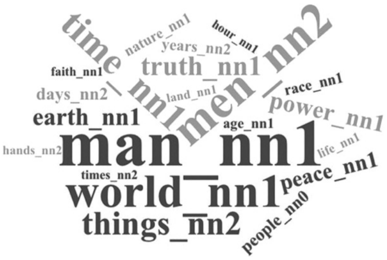


Figure 11. Word cloud of top 20 keywords (Topic 18)

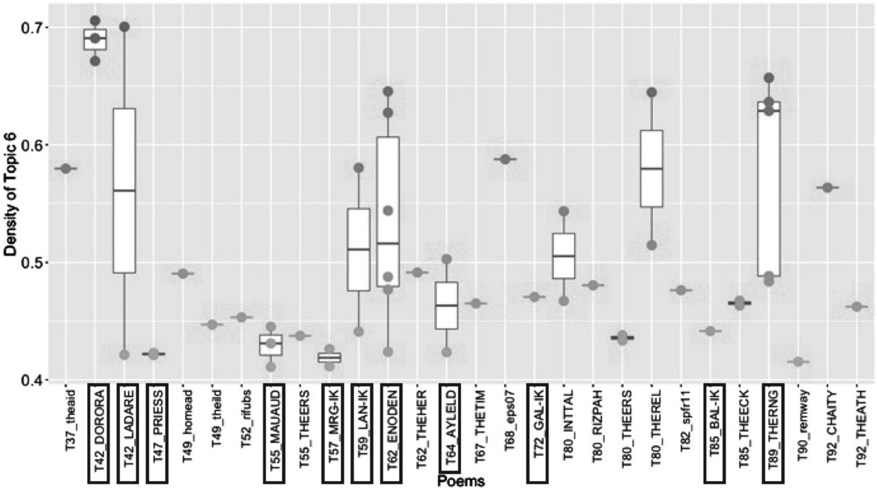


Figure 12. Density bar plot of top 50 segments of Topic 6[1]

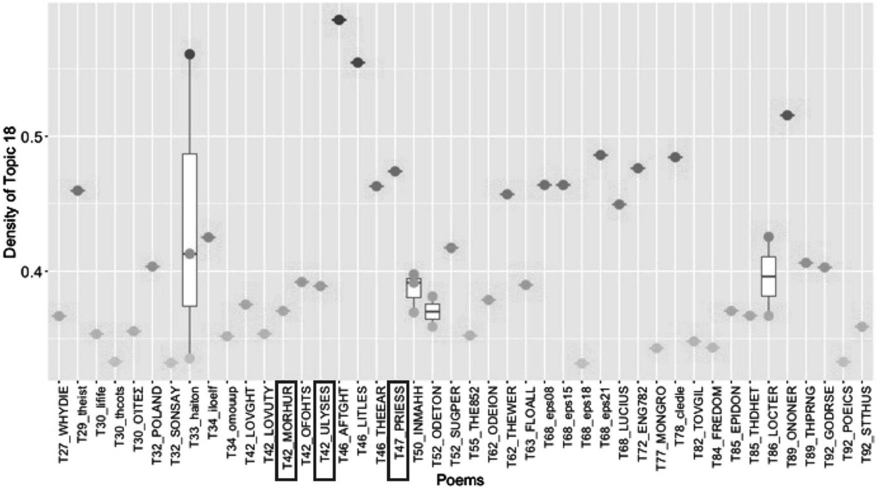


Figure 13. Density bar plot of top 50 segments of Topic 18[1]

The Princess is common to both Topic 6 and Topic 18. The protagonist of *The Princess* is a female, but this differs from the other works in these two topics. One of the poem's main themes is, to borrow the words of other scholars, 'women's role in society' (Clapp-Itnyre, 2000: 227) and 'literary and social inheritance as well as relations

Table 3. Narrative poems in top 50 segments of Topics 6 and 18

Titles of poems	Date of publication	Topic(s) appearing significantly
‘Dora’	1842	Topic 6
‘Lady Clare Vere de Vere’	1842	Topic 6
The Princess	1847	Topic 6 & Topic 18
Maud	1855	Topic 6
‘The Marriage of Geraint’	1857	Topic 6
‘Lancelot and Elaine’	1859	Topic 6
Enoch Arden	1862	Topic 6
‘Aylmer’s Field’	1864	Topic 6
‘Balin and Balan’	1885	Topic 6
‘The Ring’	1889	Topic 6
‘Mort d’Arthur’	1842	Topic 18
‘Ulysses’	1842	Topic 18

between the sexes’ (Wright, 2015: 251). More directly, it can be characterised as gender equality. While the central theme of *The Princess* is equity between the genders, the poem at the same time reflects the Victorian attitudes toward the roles of men and women (Jimbo, 2009), as the lines below show.

> Man for the field and woman for the hearth:
> Man for the sword and for the needle she:
> Man with the head and woman with the heart:
> Man to command and woman to obey;
> All else confusion.
>
> (*The Princess*, 1847: ll. 2391–2396)

In the lines above, a contrast is presented between how men and women ‘should be/should do’, and men are described as strong and wise individuals who can control women. *Enoch Arden*, which is also a significant work of Topic 6, portrays the behaviour of the main female character, Annie, toward her husband. Annie, the wife of the protagonist, Enoch, has never disobeyed her husband in the seven years of their marriage, but in the scene presented below, she opposes her husband, who is going to leave her and their family behind to go to the Far East and make a fortune:

Then first since Enoch's golden ring had girt
Her finger, Annie fought against his will:
Yet not with brawling opposition she,
But manifold entreaties, many a tear,
Many a sad kiss by day by night renew'd
(Sure that all evil would come out of it)
Besought him, supplicating, if he cared
For her or his dear children, not to go.

(*Enoch Arden*, 1862: ll. 157–164)

These lines in Tennyson's poems illustrate the position and roles of men and women in eighteenth- and nineteenth-century England: men were strong, went outside the home and existed at the centre of society, while women followed the men and stayed inside the home. The exact time period of the stories in *The Princess*, *Enoch Arden* and other poems listed in Table 3 are not necessarily mentioned directly; therefore, this paper does not assume that the historical setting is the Victorian era, but we can presume that it is safe to say these poems reflect the attitudes of society at that time and men's and women's roles during/prior to that age. Both Topic 6 and Topic 18 can be considered as suggesting the different positions of *men* and *women* and reflecting a male-dominated society in the poetic works. Additionally, Topic 6 includes the keyword *wife*, which refers to the spouse of a male, but does not include *husband*, which refers to the spouse of a female. These results indicate that Topic 6 is a male-centred topic about the home as well as Topic 18 is about male empowerment in society (especially in the context of sword-toting times/wars). Tennyson's poems tend to portray *men* as central to the home and society in/before Victorian times, and LDA detected different roles of *men* in the poems, thus Topics 6 and 18 are separated.

5. Conclusion

This study presented findings on the style of Tennyson's poetry using the LDA topic model. It showed that the latent topics behind the prominent elements of Tennyson's 603 poems can bring to light new dimensions through a balanced combination of close reading and quantitative analysis. The LDA results revealed topics that reflect the notion of immortality, one of the major themes in his body of work as well as the poet’s affinity for themes related to *sea* and *river* and to the Victorian era’s social attitudes toward *men* and *women*. Through the stylometric approach, LDA enabled us to discover new aspects of Tennyson’s poems by reading from a ‘distance’, which offers a different perspective than that of previous studies. However, this paper only investigated Tennyson’s works. To examine whether the findings are peculiar to Tennyson or not, future research will expand the analysis to compare the results with various other poets’ works, for example, William Wordsworth, John Keats and Robert Browning. All in all, through a quantitative approach, LDA revealed previously ‘unread’ parts of Tennyson’s works, although there remain the other topics not addressed in this study. More research is needed to explore these topics lest we leave them ‘unread’.

Notes

1. File titles appearing in figures comprise three parts: ‘T’ represents Tennyson; the numbers that follow reflect the last two digits of the year of publication (writing); the letters after the underscore denote the abbreviated poem title. The abbreviated poem titles in file titles can be distinguished in the following manner: works published during Tennyson’s lifetime are inscribed in upper case; poems published posthumously or unpublished are inscribed in lowercase letters. For example, T26_memyme alludes to ‘Memory [Ay me!]’ written in 1826, whereas T27_BOYOOD refers to ‘Boyhood’, which was published in 1827 when Tennyson was alive. The first line of the poem and/or Roman numerals are encased in brackets to differentiate discrete poems with identical titles. Correspondence tables of abbreviated file titles and original poem titles presented in Figures 4, 5, 8, 9, 12 and 13 (Topics 7, 17, 0, 1, 6 and 18) are available online: https://tennysondh.wordpress.com/figures-and-tables-with-their-titles/

2. On 14 May 1859, ‘Mr Peel took up Jack Tar to London; but A[lfred]. T[ennyson]. decided not to publish it’ ([Hallam Lord Tennyson], *Mat*[*erials for a Life of A.T*]. ii 217) (Ricks, 1987: 604).

Acknowledgement

This paper is a revised version of the author’s presentation at PALA 2022 *Style and Sense*(*s*), July 8, at Aix-Marseille Université. I would like to thank the chair and audience members for useful discussion at the conference. This work was supported by Japan Science and Technology Agency Support for Pioneering Research Initiated by Next Generation [grant number JPMJSP2138]. I would also like to thank two anonymous reviewers for their insightful comments and suggestions. I own responsibility for other errors.

References

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

Bloom, H. (Ed.) (1985) *Modern Critical Views: Alfred, Lord Tennyson*. New York: Chelsea House.

Bruce, H. (1917) “Tennyson and Death.” *The Sewanee Review 25*, 4: 443–456.

Clapp-Itnyre, A. (2000) “Marginalized Musical Interludes: Tennyson’s Critique of Conventionality in The Princess.” *Victorian Poetry 38*, 3: 227–248.

Delphi Poets Series (2013) *Alfred, Lord Tennyson*. East Sussex: Delphi.

Dixon, M. W. (1896) *A Primer of Tennyson*. London: Methuen & Co.

Elliot, P. (1979) “Tennyson and Spiritualism.” *Tennyson Research Bulletin 3*, 2: 89–100.

Fujita, I. (2022) “On Segment Size in Poetry Analysis Using the Latent Dirichlet Allocation Method.” *The Japanese Journal of Digital Humanities 3*, 1: 3–15.

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

Henrichs, A. (2019) “Deforming Shakespeare’s Sonnets: Topic Models as Poems Author(s).” *Criticism 61*, 3: 387–412.

Holmes, J. (2012) “‘The Poet of Science’: How Scientists Read Their Tennyson.” *Victorian Studies 54*, 4: 655–678.

Horiuchi, T. (1992) “Auditory Value in *Enoch Arden*.” *Sendai Sirayuri Junior College Bulletin 21*: 47–62.

Huang, C. (2020a) “Chugoku-no Misuterii Shousetu-ni Okeru Topikku Kaiseki-no Kokoromi (Experimental Topic Analysis on Chinese Mystery Prose Texts).” *Studies in Language and Culture Osaka University 2020*: 1–17.

Huang, C. (2020b) “Chugoku-no Misuterii Shousetu-wo Meguru Keiryouteki Bunseki: Tei Shou Sei to Ki Ba Sei-no Sakuhin-wo Chushin-ni (Quantitative Analysis of Chinese Mystery Novels: Focusing on the Works of Cheng Xiao Qing and Gui Ma Xing).” *Text mining and Digital Humanities, Graduate School of Language and Culture, Osaka University 2019*: 31–45.

Imao, Y. (2022) CasualConc (Version 3.0.2).
Available at: https://sites.google.com/site/casualconcj/casualconc/CasualConc

Iwata, T. (2015) *Topic Model*. Tokyo: Kodan-sha.

Jockers, L. M. (2014) *Text Analysis with R for Students of Literature*. Heidelberg, New York, Dordrecht, London: Springer Cham.

Jimbo, A. (2009) “Victoria-chou Shakai-no Daibensha-toshiteno Tenisun: *In Memoriam* Shuppan-no Kunou-wo Hete Eta Shousan (Tennyson, as an Opinion Leader of the Victorian Age: The Way He Won General Applause through the Publication of *In Memoriam*).” *The Bulletin of the Graduate School of Education of Waseda University 17*, 1: 355–364.

Kabata, T. (2001) “‘Nothing’ in Tennyson’s Poems.” *Bulletin of Yonezawa Women’s College 36*: 21–31.

Kabata, T. (2007) “‘Sun’ in Tennyson’s Poems.” *Bulletin of Yonezawa Women’s College 42*: 9–20.

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*: 79–99.

Kuroda, A. (2017) “Quantitative Analysis of Literary Works: Novels of Sir Arthur Conan Doyle.” *Text mining and Digital Humanities, Graduate School of Language and Culture, Osaka University 2016*: 23–41.

McCallum, A. K. (2002) MALLET: A MAchine Learning for LanguagE Toolkit.
Available at: http://mallet.cs.umass.edu.

Matsukawa, H., M. Oyama, C. Negishi, and Y. Arai (2018) “Analysis of the Free Descriptions Obtained through Course Evaluation Questionnaires Using Topic Modeling.” *Japan Journal of Educational Technology 41*, 3: 233–244.

Moretti, F. (2013) *Distant Reading*. London: Verso.

Nakamura, M. (1967) “Tennyson and his Christianity.” *The Journal of Otemae Women’s University 1*: 76–87.

Navarro-Colorado, B. (2018) “On Poetic Topic Modeling: Extracting Themes and Motifs from a Corpus of Spanish Poetry.” *Frontiers in Digital Humanities 20*: 5–15.

Nishimae, Y. (1979) *Tenisun Kenkyu* (*Tennyson Research*). Tokyo: Chuo-Shuppan.

Nishimae, Y. (2000) *Tenisun-no Gengo Geijutsu* (*Tennyson’s Linguistic Art*). Tokyo: Kaibunsya.

Noguchi, T. (2011) “Tennyson’s Ulysses in the Dramatic Monologue: The Dramatization of Visions Through the Wreck.” *Hokusei Review, the School of Humanities 48*, 2: 1–11.

Okabe, M. (2019) “*Thou* and *You* in Emily Dickinson’s Poems Using Topic Modeling: Reconsideration of Interjections.” *Proceedings of Japanese Association for Digital Humanities Conference 2019*: 125–131.

Okazawa, T. (1969) “A Study of Tennyson’s Predilection for Waters.” *Bulletin of Chofu Gakuen Women’s Junior College 2*: 1–12.

Onodera, D., L. Huang, and M. Yoshioka (2016) “Classification of New Article by Using Facet-Biased Topic Model and Distance Metric Learning.” *The 30th Annual Conference of the Japanese Society for Artificial Intelligence 2016*: 1–4.

Oxford English Dictionary (online) https://www.oed.com Last accessed date: October 13th, 2022.

Perrine, L. (1966) "When Does Hope Mean Doubt?: The Tone of 'Crossing the Bar'." *Victorian Poetry 4*, 2: 127–131.

Plamondon, M. R. (2005) "Computer-Assisted Phonetic Analysis of English Poetry: A Preliminary Case Study of Browning and Tennyson." *TEXT Technology*: 153–175.

Rayson, P. (2009) Wmatrix: A Web-Based Corpus Processing Environment, Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

Rhody, M. L. (2012) "Topic Modeling and Figurative Language." *CUNY Academic Works 2012*: 19–35.

Ricks, C. (Ed.) (1969) *The Poems of Tennyson*. London: Longman.

Ricks, C. (Ed.) (1987) *The Poems of Tennyson*. vol. I–III. (Second edition) London: Longman.

Shaw, W. D. (1973) *Tennyson's Style*. Ithaca and London: Cornell University Press.

Shaw, W. D. (1976) "Tennyson's 'Tithonus' and the Problem of Mortality." *Philological Quarterly 52*, 2: 274–285.

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*, Doshisha University.

Tabata, T. (2018) "Mapping Dickens's Novels in a Network of Words, Topics, and Texts." *Text mining and Digital Humanities, Graduate School of Language and Culture, Osaka University 2016*: 51–60.

Tabata, T. (2020) "Latent Topics in British Classic Fiction: Using LDA to Classify Texts into Meaningful Groups." *Text mining and Digital Humanities, Graduate School of Language and Culture, Osaka University 2019*: 47–58.

Tennyson, H. (1897) *Alfred Lord Tennyson: A Memoir by his Son*. vol. I–III. London: Macmillan.

Thomas, J. (2019) *Tennyson Echoing Wordsworth*. Edinburgh: Edinburgh University Press.

Wright, J. (2015) "*The Princess* and the Bee." *The Cambridge Quarterly 44*, 3: 251–273.

（藤田　郁　大阪大学院）

「論文」

Measuring Similarities Within Word Families: A Word-embedding Approach Using word2vec

Satoru UCHIDA and Mitsuhiro MORITA

Abstract

The word family is a useful concept to determine the lexical aspects of English learners and has been widely used in vocabulary studies. However, it has been criticized, especially because elementary foreign language learners do not have a full command of its derivational operations. In addition, it remains unclear as to which member in a word family is challenging for the learners. This study examines the similarities between each member of the word family by using word2vec, a widely used natural language processing application. Based on the similarity scores between the word forms generated by the application using 7,540 pairs of words created from the CEFR-J wordlist and BNC/COCA family lists, this study argues that teachers and learners must especially focus on word families with low similarity scores. Furthermore, these results are useful for determining the difficulty level of affixes and discovering specific word forms that require special treatment in the classroom.

1. Introduction

A word family is a group of words with a word base or stem. For example, “kindly,” “kindness,” and “unkind” are grouped into one word family with “kind” as the common base. The concept of word family is important in English vocabulary studies, especially with regard to vocabulary assessment and text coverage.

In the context of English as a foreign language (EFL) classrooms, it should be considered that the members in a word family do not always exhibit the same difficulty level. For instance, it would be easy for several learners to infer the meaning of “flatten” from the meaning of “flat” with the knowledge of “en” as a verb suffix, but linking the meaning of “flat” and “flatly” would be more challenging, especially for a novice learner who has just learned such phrases as "a flat stone." This implies that teachers and learners must be aware that each member in a word family has an individuality. In fact, some studies have revealed the issues with using word families as the measuring unit for assessing the learners' vocabulary size (Gardner, 2007; Kremmel, 2016). However, no attempts have been made to identify which words in each word family actually cause problems for learners.

The present study aims to evaluate the similarities among each member in a word family using word2vec (Mikolov et al., 2013), a powerful and influential application in natural language processing (NLP) that enables the assignment of numbers (vectors) to words; these can then be observed as a representation of word meanings. Subsequently, we can calculate how closely each word is related using cosine similarity scores. The current study hypothesizes that this score can be used to measure the relatedness and learnability of words within a word family, which eventually reveals peculiar members that need special attention in English education.

2. Literature Review

2.1 Word family and vocabulary learning

Word families have been used as counting units in vocabulary research, especially vocabulary knowledge assessment and text coverage studies. Popular vocabulary assessment measures have adapted word family counts, such as the Vocabulary Levels Test (Nation, 1983) and the Vocabulary Size Test (Nation & Beglar, 2007), among others. Nation's (2006) influential text coverage study indicated that for 98% coverage, the most frequent 8,000 to 9,000 word families were necessary for written discourse, and the most frequent 6,000 to 7,000 word families were essential for spoken discourse. Nurmukhamedov and Webb (2019) reported that many text coverage studies adapted word families as counting units owing to the development of corpus analysis tools, which enabled researchers to create corpus-based word lists, such as the BNC/COCA (British National Corpus/Corpus of Contemporary American English) word family lists, and computerized text analysis tools based on word lists, such as Range. These studies all use a word family count with the expectation that "once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort" (Bauer & Nation, 1993: p. 253).

However, literature has also challenged the idea of the word family as a counting unit (Gardner, 2007; Kremmel, 2016; McLean, 2018; Reynolds, 2013; Stoeckel et al., 2021). Some empirical studies have indicated that learners of EFL experienced difficulty in learning derived words and provided evidence to support the challenge. Research with Japanese learners of English has shown that they had insufficient derivational knowledge (Schmitt & Meara,1997; Mochizuki & Aizawa, 2000; McLean, 2018). Among those studies, McLean (2018) examined 279 university-level Japanese learners of English regarding their knowledge of inflections and derivations. Participants were asked to write a Japanese translation for a target item. The accuracy rate for inflections was 98% when the participants knew the bases, whereas it was 54% for the derivations. Moreover, studies with Thai and Austrian learners of English revealed that knowing the base words did not guarantee that the learners would be aware of their derivations (Ward & Chunenjudaeng, 2009; Kremmel & Schmitt, 2016). Based on these studies, Brown et al. (2020) claimed that the lemma (the baseword and inflected forms of a word of a particular part of speech) or flemma (the base form and inflected forms of a word, regardless of part of speech) is the more appropriate counting unit for second language (L2) English learners than word family due to their limited derivational knowledge. Although Laufer (2021) argued strongly against the claim that word family is not suitable for the counting unit, it is proposed that the Nuclear Family List developed by Cobb and Laufer (2021) consisting of frequently used word family members, should be used for novice and intermediate learners to expose them to useful derived words. Thus, it is clear that some effort is required for L2 learners, at least in their early stages, to learn the members of a word family.

The current issue examines how English instruction can help learners effectively expand the members of word families. One way to accomplish this is to increase the learners' exposure to derived words in teaching and learning materials. However, it has been noted that this method provides limited input for the learners. Laufer and Cobb (2020) examined the frequency of prefixes and suffixes in graded readers as well as a limited number of academic articles, news articles, and novels; their results indicated that only a few prefixes and suffixes were required to read the texts. The graded readers examined in the study required "-ly" and "-y" suffixes to understand 98% of the text. Morita et al. (2019) investigated prefixes and suffixes in junior high school English textbooks in Japan to reveal that only limited types and tokens of prefixes and suffixes were used. A similar result was found in high school English textbooks (Morita et al., 2021). While first language (L1) studies demonstrated that elementary school students encounter far more words with prefixes and suffixes (Anglin et al., 1993; Nagy & Anderson, 1984), graded readers and textbooks may not be sufficient for learners to grasp derivational words.

Another way to foster the learners' mastery of derivational words is to provide explicit instruction. While individual studies for explicit morphological instruction showed mixed results (e.g., Sritulanon, 2013 for not effective; Lin, 2019 for effective; Ross & Berwick, 1991 for effective in a limited domain), recent meta-analysis studies have found that explicit instruction regarding derivational affixes benefits both L1 and second language L2 English learners (Goodwin, 2016; Goodwin & Ahn, 2010; Kirby & Bowers, 2017). However, it is unclear which members in a word family are suitable for learning and teaching derivational forms. Therefore, the current study aims to effectively bridge this gap.

2.2 The present study

Extant literature has revealed that the exposure to affixes through learning materials is limited, although certain prefixes and suffixes occur more frequently. While it remains ambiguous as to which words in each word family are difficult or easy for learners, clearly some affixes are easier to master than others. One of the underlying factors is the combination of the base and affix. Specifically, "player" may be simple to learn, but the meaning of "sitter" is not directly drawn from "sit." This fact suggests that the difficulty level of a derivation should be judged word by word. Therefore, the present study attempts to prove the usefulness of word2vec, a widely used natural language processing (NLP) application, to reveal the derivations that need special treatment in teaching and learning English. It is anticipated that word forms that display unique behavior have lower similarity scores between the base form. For example, the usage of "lastly" differs from its base form "last" in that the former is used as a list marker while the latter can be used as either a verb or an adjective. It is expected that we may be able to identify word forms that require special treatment in education by observing the similarity scores.

3. Methodology

3.1 word2vec

According to the distributional hypothesis proposed by Harris (1954), words that denote similar meanings occur in similar contexts (see Sahlgren, 2008 for a detailed discussion). As a recent NLP technique, word embedding relies on this theory to map the word meanings to a set of numbers (vectors) using contextual information. A simplified model is described as follows to explain this process using the collocational information of the sample words.

Six nouns are selected here, one of which is masked for the purpose of demonstration (“apple,” “car,” “cat,” “dog,” “XXX” (masked), and “pencil”). Table 1 displays the frequencies of the verbs and adjectives (“buy,” “drive,” “eat,” “fresh,” “peel,” “sharpen,” and “stray”) with the target nouns taken from the COCA using the following expressions: “[*verb*] [a] [*noun*]” and “[a] [*adjective*] [*noun*],” which allow us to include the inflected forms, such as “eats” and “ate.” If we need to determine “XXX” in this table, one possible approach involves comparing the frequencies of the collocations to discover a word with a tendency similar to “XXX,” presuming that such a word has something in common with the masked word. This simple method of using contextual (collocational) word information is based on the distributional hypothesis.

Table 1. The Sample Words’ Verbal and Adjectival Collocations

	buy	drive	eat	fresh	peel	sharpen	stray
apple	68	0	140	112	19	0	0
car	1,118	1,428	1	5	0	0	1
cat	16	1	15	3	0	0	297
dog	93	1	32	9	0	0	533
XXX	9	2	28	245	41	0	1
pencil	3	1	3	4	0	17	1

However, it is difficult to identify the most similar word at a glance from the given table of raw frequencies. To mathematically calculate the distances between each word, the cosine similarity score is beneficial; it can be calculated using the following formula where x_i and y_i denote the frequencies of each collocation of the target words:

$$\cos(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

This score ranges from -1 to 1 and takes a value of 1 if all the numbers are identical. Table 2 displays the matrix of the cosine similarity scores between the sample nouns.

Table 2. Cosine Similarity Scores Between the Sample Nouns

	apple	car	cat	dog	XXX	pencil
apple	1.000	0.220	0.061	0.113	**0.679**	0.309
car		1.000	0.036	0.108	0.031	0.147
cat			1.000	0.993	0.021	0.075
dog				1.000	0.033	0.097
XXX					1.000	0.243
pencil						1.000

The highest score is 0.993 between “dog” and “cat,” given the simple fact that these two nouns are animals, and the others are not; naturally, the collocations of “dog” and “cat” are fairly similar. In the “XXX” column, the most similar word is “apple” (0.679). A natural guess is that “XXX” also denotes a type of food, or more specifically, a fruit. Actually, the answer is “orange.” Note that the meanings of “apple” and “orange” can be denoted as a set of numbers—or [68, 0, 140, 112, 19, 0, 0] and [9, 2, 28, 245, 41, 0, 1], respectively—which enables a calculation of the distance (similarity score) between them.

The word2vec application is based on this framework with mathematical sophistication. It was developed by Mikolov et al. (2013) and has been widely used in the NLP field, but has been scarcely utilized in the fields of linguistics and applied linguistics. Word2vec employs a shallow neural network model to efficiently learn the vector representations of words, although a detailed explanation is beyond the scope of this paper (see Goldberg & Levy, 2014 for a detailed explanation). If we use the same previously explained approach with the entire COCA, the list of collocations continues endlessly up to the number of the types of words. The word2vec application automatically groups the collocations into certain dimensions, typically 200 to 500, to define the "meaning" of the words with a list of numbers.

Using the full-text data from the COCA (1990–2015; approximately 600 million words), we created a word2vec model with a *gensim* library in Python. To vectorize per sentence, the *sent_tokenize* function in the *nltk* library was utilized to separate the text into sentences, which were then converted to lowercase with the *LineSentence* function used for modeling.[1]

For example, this model generates a vector for "orange" of [-0.782, 1.389, -1.732, …, -0.234], or 300 numbers in total; if we calculate the top five similar words in terms of the cosine similarity score, we get "yellow" (0.682), "tangerine" (0.620), "peach" (0.606), "blue" (0.602), and "red" (0.600). As anticipated, these words relate to either colors or fruits. It should be noted that this pseudo-representation of word meanings reflects both semantic and syntactic characters. In other words, low similarity scores indicate semantic as well as syntactic differences between the word pairs.

One note to be added here is that the list may contain what are considered antonyms. For example, the closest word for "increase" in our model is "decrease" (0.878). This is a natural result considering that these two words can appear in extremely similar contexts. These can even be used interchangeably, in such sentences as "The number of students in the university *increased* [decreased] by 5% last year." This is at times perceived as a disadvantage of word embedding but is rather advantageous in this study. Antonyms typically belong to the same word family, including "like" and "dislike," "known" and "unknown," and "useful" and "useless." On the one hand, if the contexts are similar enough, word2vec will assign high similarity scores to the antonym, which may imply that there is no need for special treatment in the classroom. On the other hand, if the score for the relationship between the base word and its antonym is low, this suggests the need for special attention.

3.2 Word family list

One of the most extensive word family lists is Paul Nation's BNC/COCA family lists, a modified version of which is available as "BNC/COCA family lists + extras" (ver. 2.00)[2] from AntLab (managed by Laurence Anthony). This list contains headwords with word family members with the following examples:

- A
    - AN
- ABLE
    - ABILITIES
    - ABILITY
    - ABLER
    - ABLEST
    - ABLY
    - INABILITY
    - UNABLE
- ABOUT
- ABOVE
- ABSOLUTE
    - ABSOLUTELY
    - ABSOLUTES
    - ABSOLUTISM
    - ABSOLUTIST
    - ABSOLUTISTS

The list includes inflections, such as “abilities” as the plural form of “ability,” in addition to derivations; for instance, “ability,” “inability,” and “unable” are derived from “able.” It has 50,890 headwords and 105,476 forms, including each base form.

3.3 The CEFR-J wordlist

CEFR-J wordlist[3] (ver. 1.6), which is widely used in Japan, is employed as a testing vocabulary set to focus on words that are useful for English learners. The advantage of this list is that it classifies words into the CEFR levels (A1, A2, B1, and B2), which comprise a common structure to assess language ability. It indicates which word is more difficult for the learners to master, and thus, is a possible indicator of the difficulty of the word forms in a word family. Although the CEFR-J wordlist has different levels of the same surface form (e.g., sentence [noun]): A1, sentence [verb]: B2), for the sake of simplicity, the level of the highest one on the wordlist was chosen here.

The CEFR-J wordlist contains 7,801 words; we excluded compounds (e.g., “bus stop” and “each other”) and words without any inflections or derivations (e.g., “about” and “above”). Furthermore, we ignored the low-frequency British forms (e.g., “industrialise” and “familiarise”) as well as the words that appear less than 16 times in our dataset (e.g., “narcissistic” and “aubergine”) to ultimately yield a list of 6,290 words.

3.4 Dataset

We used the word family list to assign family members to each headword in the selected list (e.g., “ability,” “abilities,” and “inability,” among others). This resulted in a list of 17,206 pairs, such as “able–ability,” “able–inability,” and “absolute–absolutist,” with an average of 3.95 pairs per headword. As these pairs include rare word forms that can be considered unimportant to learners, the top 20,000 most frequent word forms in the COCA are chosen. This process excludes such word pairs as “absolute–absolutist,” “bottom–bottoming,” and “computer–computationally.” Finally, 7,540 unique pairs were incorporated into our target dataset; Figure 1 illustrates the process of creating our dataset.

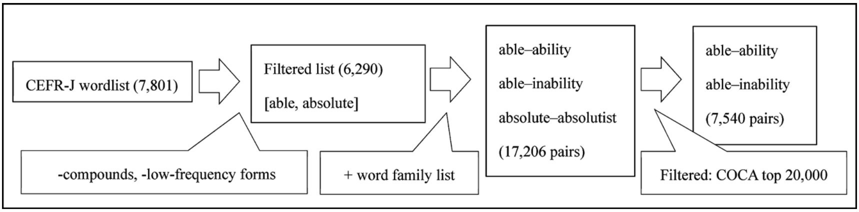


Figure 1. Creating the Target Dataset

4. Finding word pairs that deserve special attention

We hypothesized that words with a low similarity score are difficult to learn and hence require special attention in teaching and learning. To consider the CEFR levels of each word form, we selected 1,543 word pairs out of the 7,540 pairs in which both words were assigned a CEFR level for the current analysis. Table 3 displays some random examples of word pairs with low average similarity scores whose similarity score (represented as cos. in the table) is equal to or below 0.1.

The lowest-scoring word pair in this table is “detect” and “detective.” These two forms seem to require special attention in the classroom. The word “detective” refers to a person or a police officer who investigates crimes; this word is typically taught on a different occasion from the verb “detect,” which is primarily used in the context of experiments or machines (sensors). Thus, learners may consider these as different vocabulary words and may impose extra cognitive costs, although linking these words

Table 3. Examples of Word Pairs with a Low Similarity Score

word form 1	word form 2	cos.	word form 1	word form 2	cos.
detect (B2)	detective (B1)	-0.061	conduct (B1)	conductor (B2)	0.039
expect (A2)	unexpectedly (B1)	-0.035	suppose (B1)	supposedly (B1)	0.043
time (A1)	timeless (B2)	-0.025	exhaust (B2)	exhausting (B2)	0.044
vary (B1)	invariably (B2)	-0.025	remark (B2)	remarkably (B2)	0.053
double (A2)	doubly (B2)	-0.025	late (A1)	lately (B1)	0.054
over (A2)	overly (B2)	-0.014	character (A1)	characterize (B1)	0.063
count (B1)	countless (B1)	-0.001	end (A1)	endlessly (B2)	0.067
drama (A1)	dramatically (B2)	0.009	point (A1)	pointless (B2)	0.077
last (B1)	lastly (B2)	0.011	govern (B1)	governor (B1)	0.080
suit (A2)	suitable (A2)	0.018	ready (A1)	readily (B2)	0.080
time (A1)	timely (B1)	0.019	second (A1)	secondly (B2)	0.085
lightly (B1)	lighten (B2)	0.025	special (A1)	specialize (B1)	0.093
belong (A2)	belongings (B2)	0.028	mean (A2)	meaningful (B1)	0.094
total (B1)	totally (B1)	0.038	bear (A1)	unbearable (B2)	0.096
pave (B2)	pavement (B2)	0.039	remark (B2)	remarkable (B1)	0.100

using the word family concept would help them efficiently understand and remember the word meanings. Some other word pairs in Table 3 are also remarkable. For example, “last” and “lastly” as well as “suit” and “suitable” are clear cases of caution for teachers and learners. Specifically, the base forms are polysemous, and some do not explicitly relate to the meanings of the derived forms—and hence, deserve special attention. Other cases to note are “belong” (A2) and “belongings” (B2), “total” (B1) and “totally” (B1), and “character” (A1) and “characterize” (B1). These cases all demonstrate the usefulness of the word2vec similarity scores in discovering the word families that demand special care in an English educational context. It should be noted that the results reflect not only the semantic differences of each word but also their syntactic differences. For instance, there is a clear syntactic difference between “last” and “lastly,” especially since the latter functions as a sentential adverb, and the environment in which they appear is substantially distinct.

If low-scoring word pairs require special attention, it can be assumed that one of the word forms is at a higher level. In other words, it can be predicted that the word pairs with different CEFR levels (e.g., A2–B1) will have lower similarity scores than the word pairs with the same CEFR level (e.g., B1–B1). Around 537 pairs (approx. 35%) out of 1,543 have the same CEFR level [e.g., "manage" (A2) and "manager" (A2)] whereas others (1,006 pairs) have different levels [e.g., "know" (A1) and "unknown" (A2)].

We conducted the Welch two-sample *t*-test (R ver. 4.1.0) with the similarity scores of the two groups. The findings revealed that the words with the same CEFR level [$N = 537$, $M = 0.42$, $SD = 0.18$] have significantly higher similarity score (t (1034.7) = 6.10, $p < .001$, $d = 0.33$ [0.23, 0.44]) than the words with different levels ($N = 1{,}006$, $M = 0.36$, $SD = 0.17$). In other words, the difference in the CEFR levels significantly contributes to the variance in the scores of each group. It should be noted that there are cases where each word needs special attention even when the words share the same CEFR level, and these cautious words can be located by using the similarity score. For example, in Table 3, both "suit" and "suitable" are labeled as A2, but they may be worth extra attention presumably due to the polysemy of the verb "suit." Thus, it can be concluded that the word2vec scores serve as a suitable indicator of word difficulty within a word family, particularly in terms of the CEFR level, and are useful in finding words of caution for teachers and learners.

5. Applications of the research results

5.1 Finding additional forms to be listed in the wordlist

The previous section has proved the usefulness of word2vec in identifying words that need special attention. However, we have confined our scope to cases of word forms listed in the CEFR-J wordlist. As an application of our methodology, this subsection attempts to discover word forms that are not currently included there but are noteworthy and could be added in a future edition of the wordlist. We are aware that each wordlist has its own policy for the selection of the words. For example, a wordlist may only include those words with basic derivations assuming that the complex forms (e.g., "unkindness") are rare, and their meanings could be deduced from its more basic form. However, even certain simple derivations could be noteworthy, and the following section is an attempt to find those word forms.

We used the CEFR-J wordlist as a base list to investigate each headword's family members from the BNC/COCA family lists. Subsequently, the similarity scores were calculated against its base form for the derived forms that are not included in the CEFR-J wordlist. For example, "acceptability" and "accepted" are not part of the CEFR-J wordlist, and therefore, we calculated their similarity scores with the base form "accept" (0.128 and 0.520, respectively). Given the 5,997 pairs (7,540–1,543) employed for this experiment, low similarity scores suggest that the word form behaves differently from the base form, hence requiring special attention. Table 4 lists the word forms with the lowest similarity scores against each base form.

Table 4. The Lowest Five Word Forms in the Similarity Score Against the Base Forms

Rank	Base	CEFR	COCA Rank	Form	CEFR	COCA Rank	cos.
1	decide	A2	1,663	decidedly	#N/A	11,057	-0.152
2	center	A2	8,194	centrist	#N/A	18,208	-0.110
3	correct	A1	1,876	correctional	#N/A	15,699	-0.107
4	mark	A2	3,728	markedly	#N/A	13,559	-0.101
5	offend	B2	14,929	offenders	#N/A	7,520	-0.091

This list includes the rare word forms (e.g., "centrist" and "correctional") and an inflected form (e.g., "offenders") but has simultaneously derived forms that deserve attention (e.g., "decidedly" and "markedly") with meanings and syntactic behaviors that clearly differ from the base form. Table 5 lists the words that were manually selected from our experimental results.

These word forms, which have low similarity scores and are not currently included in the CEFR-J wordlist, can be considered as irregular members within a word family, and hence, deserve special treatment in classrooms. The results indicate that certain base forms (e.g., "elevate" and "refine") are not included in the list. Furthermore, some frequent and important word forms (e.g., "notably" and "namely") are missing. Therefore, our approach is successful in finding the word forms that deserve consideration for future addition to the list.

Table 5. Examples of Word Forms with Low Similarity Scores

Base	CEFR	COCA Rank	Forms	CEFR	COCA Rank	cos.
elevate	#N/A	16,855	elevator	A2	5,211	-0.075
strike	A2	3,849	strikingly	#N/A	15,493	-0.029
note	A1	1,604	notably	#N/A	6,434	-0.017
react	B1	5,118	reactor	#N/A	9,807	-0.004
flat	B1	2,087	flatly	#N/A	15,426	-0.004
period	A1	669	periodically	#N/A	9,565	0.005
name	A1	376	namely	#N/A	6,827	0.008
man	A1	136	unmanned	#N/A	18,315	0.011
refine	#N/A	15,701	refinery	B2	17,599	0.012
man	A1	136	manned	#N/A	19,537	0.015

5.2 Difficulties of affixes

The similarity score can be calculated for each word form against its base, and it is possible to estimate the difficulties of affixes if we compute the average scores of the words with a specific affix. For instance, by comparing the average score of “kind–unkind,” “conscious–unconscious,” and “aware–unaware” against the average of “take–mistake,” “understand–misunderstand,” and “use–misuse”, it would be clarified which (“un” or “mis”) affix should receive more attention in teaching and learning.

For this purpose, we use the 1,543 word pairs with CEFR levels employed in the previous section. Based on the affix levels proposed by Bauer and Nation (1993), each pair is grouped into major affix patterns, such as “-ly,” “un-,” and “-ment.” For example, the pairs “total–totally” and “kind–kindly” can be grouped as “-ly” pairs because both their derivation forms have the “-ly”’ ending. Table 6 displays the affix levels, which are based on the affixes’ productivity, predictability, and regularity, among other traits. It is assumed that elementary learners are only aware of affixes at the lower levels, but advanced learners know the higher-level affixes as well.

Consequently, 1,333 pairs out of 1,543 (86.4%) are given an affix classification while such irregular pairs as “we–ourselves,” “who–whoever,” and “that–those” are not classified. Each pair’s scores were calculated based on their derivation and base forms; then, they were averaged within each group. For example, the “-ment” group has 43 pairs—such as “announce–announcement” (0.403), “measure–measurement” (0.585),

Table 6. Affix Levels (Bauer & Nation, 1993)

Level	Description
	Affixes
1	Each form is a different word
2	Inflectional suffixes
3	The most frequent and regular derivational affixes
	-able, -er, -ish, -less, -ly [adv.][1], -ness, -th [ordinal number], -y [adj.], non-, un- [antonym] (all with restricted uses)
4	Frequent, orthographically regular affixes
	-al [adj.], -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in- (all with restricted uses)
5	Regular but infrequent affixes
	-age, -al [noun], -ally, -an, -ance, -ant, -ary [adj.], -atory, -dom, -eer, -en [adj.], -en [verb], -ence, -ent, -ery, -ese, -esque, -ette, -hood, -i, -ian, -ite, -let, -ling, -ly [adj.], -most, -ory, -ship, -ward, -ways, -wise, ante-, anti-, arch-, bi-, circum-, counter-, en-, ex-, fore-, hyper-, inter-, mid-, mis-, neo-, post-, pro-, semi-, sub-, un- [reverse]
6	Frequent but irregular affixes
	-(ate+)able[2], -ee, -ic, -ify, -ion, -ist [adding to unexplained consonant], -ition, -ive, -th, -y [noun], pre-, re-
7	Classical roots and affixes

Note. [1]Brackets [ ] after some affixes include the part of speech, meaning, or other information produced by attaching the affix. [2](*-ate*) before *-able* means that *-ate* is typically omitted, such as *permeable*.

and "pay–payment" (0.506)—with an average score of the pairs in this group being 0.313 (figures in parentheses indicate each similarity score). Some words were grouped into multiple categories, such as "will–willingness" ("-ing," "-ness"), "create–creativity" ("-tive," "-ity"), and "vary–invariably" ("in-," "-able," "-ly"). However, these are excluded from the calculation to avoid mixed effects. It is assumed that the average score represents the affix's difficulty. Table 7 displays the highest 10 and lowest 10 groups of affixes with more than four words.

As the table indicates, the following inflectional affixes have relatively high scores: "-s," such as "sport–sports" (0.702) and "thank–thanks" (0.695); "-ing," such as "wrap–wrapping" (0.709) and "feel–feeling" (0.656); and "-ed," such as "attach–attached" (0.695) and "worry–worried" (0.689). Additionally, various morphemes that

Table 7. Average Scores of Each Affix Group with More Than Four Words

	Highest 10			Lowest 10		
No.	Morpheme (Level)	Average	Count	Morpheme (Level)	Average	Count
1	-s (2)	0.509	15	-less (3)	0.141	18
2	in- (4)	0.498	15	-or (NA)	0.213	23
3	-ing (2)	0.484	127	-able (3)	0.272	22
4	-an (5)	0.463	6	-ence (5)	0.307	25
5	-ism (4)	0.460	6	-ment (4)	0.313	43
6	-ic (6)	0.442	22	-ive (6)	0.315	25
7	un- (3)	0.431	29	-er (3)	0.318	105
8	-ed (2)	0.429	80	-age (5)	0.344	7
9	-ist (6)	0.429	17	-ness (3)	0.349	23
10	-ship (5)	0.426	6	-ary (5)	0.358	5

create antonyms have high scores: “in-,” such as “accurate–inaccurate” (0.672), “expensive–inexpensive” (0.642), and “effective–ineffective” (0.624); and “un-,” such as “aware–unaware” (0.729), “familiar–unfamiliar” (0.679), and “happy–unhappy” (0.632). This is likely because they do not change the part of speech and are used in similar contexts. As morphemes with high scores are relatively easy to master, teachers may assume that learners do not need morphological instructions for them. However, even when an affix group can be considered as being straightforward for learners, some word forms deserve special attention. The word2vec results make it possible to list such cases. For example, “-s,” “-ing,” and “-ed” are simple inflectional endings, but such low-scoring word pairs as “mean–means” (0.392), “custom–customs” (0.285), “miss–missing” (0.119), “concern–concerning” (0.184), “puzzle–puzzled” (0.147), and “mark–marked” (0.173) should be carefully addressed in the classroom. Moreover, it should be noted that a few of the level 3 affixes—such as “-less,” “-able,” “-er,” and “-ness”—are contained in the lowest 10 list. Although these affixes are productive and predictable, their low scores imply the necessity for extra instruction. For instance, such word pairs as “point–pointless” (0.077), “need–needless” (0.095), and “hope–hopeless” (0.136) have low similarity scores, which implies that their usage and meaning of derivation highly differ from their base forms. Additionally, “-er,” which is one of the simplest suffixes, is not free from irregularity; its various word pairs—such as “begin–beginner” (0.023), “train–trainer” (0.140), “stick–sticker” (0.150), “deal–dealer” (0.184), and “hold–holder” (0.211)—may require special treatment.

6. Conclusion

This study examined the usefulness of an NLP application called word2vec in analyzing word families. We employed the similarity scores generated by the application to assess the similarities among the word forms in each word family, uncovering cases that deserve special attention in teaching and learning English. In other words, the low similarity scores between the derivation and base forms clearly indicated the necessity for supplemental instruction on the word pair in the classroom context. Additionally, the results identified some word forms that could be included in a vocabulary list. Reorganizing the findings based on the affix groups revealed each affix’s difficulty level as well as the information on remarkable words within the group. These results reveal the usefulness of word2vec and potential possibilities for collaborations between the NLP and the English education fields.

The limitation of this study is that it discussed the similarity scores in relation to word difficulty but did not actually experiment with English language learners as subjects. Although there were statistically significant differences based on the CEFR levels, a more detailed analysis would be needed to objectively determine the difficulty level of each word form using the word pairs presented in this study. Furthermore, the focus of this study was on finding words in a word family that need special consideration in class and not on how to teach them. Future research might include a more qualitative analysis of the words with low similarity scores, including how they should be treated for different levels of learners.

Notes

[1] The parameters for the model were: size = 300, window = 5, min_count = 3, and iter = 5, and the other parameters were set to default.

[2] http://www.laurenceanthony.net/resources/wordlists/bnc_coca_cleaned_ver_002_20141015.zip

[3] The CEFR-J Wordlist Version 1.6. Compiled by Yukio Tono, Tokyo University of Foreign Studies. Retrieved from http://www.cefr-j.org/download.html in June, 2020.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22H00677.

References

Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development, 58*(10, serial no. 238). https://doi.org/10.2307/1166112

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Brown, D., Stoeckel, T., McLean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, *43*(3) 1-7. https://doi.org/10.1093/applin/amaa061

Cobb, T., & Laufer, B. (2021). The Nuclear Word Family List: A list of the most frequent family members, including base and affixed words. *Language Learning*, *71*(3), 834–871. https://doi.org/10.1111/lang.12452

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241–265. https://doi.org/10.1093/applin/amm010

Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Goodwin, A. P. (2016). Effectiveness of word solving: Integrating morphological problem-solving within comprehension instruction for middle school students. *Reading and Writing*, *29*(1), 91–116. https://doi.org/10.1007/s11145-015-9581-0

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, *60*(2), 183–208. https://doi.org/10.1007/s11881-010-0041-x

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Kirby, J. R., & Bowers, P. N. (2017). Morphological instruction and literacy: Binding phonological, orthographic, and semantic features of words. In. *Studies in Written Language and Literacy* D. L. C. Cain & R. K. Parrila (Eds.), (437–462). https://doi.org/10.1075/swll.15.24kir.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987. https://doi.org/10.1002/tesq.329

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary Test Scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Laufer, B. (2021). Lemmas, flemmas, word families and common sense. *Studies in Second*

Language Acquisition, *43*(5), 965–968. https://doi.org/10.1017/S0272263121000656

Laufer, B., & Cobb, T. (2020). How much knowledge of derived words is needed for reading? *Applied Linguistics*, *41*(6), 971–998. https://doi.org/10.1093/applin/amz051

Lin, M.-F. (2019). Developing EFL learners' morphological awareness: Instructional effect, teachability of affixes, and learners' perception. *International Review of Applied Linguistics in Language Teaching*, *57*(3), 289–325. https://doi.org/10.1515/iral-2015-0081

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amw050

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, *28*(2), 291–304. https://doi.org/10.1016/S0346-251X(00)00013-0

Morita, M., Uchida, S., & Takahashi, Y. (2019). The frequency of affixes and affixed words in Japanese junior high school textbooks: A corpus study. *ARELE: Annual Review of English Language Education in Japan*, 30, 129-143.

Morita, M., Uchida, S., & Takahashi, Y. (2021). The frequency of affixes and affixed words in Japanese senior high school English textbooks: A corpus Study. *ARELE: Annual Review of English Language Education in Japan*, 32, 81-95.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, *19*(3), 304–330. https://doi.org/10.2307/747823

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, P. (1983). Teaching and testing vocabulary, *Guidelines*, *5*(1), 12–25.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *Language Teacher*, *31*, 9–13. https://jalt-publications.org/tlt/issues/2007-07_31.7.

Nurmukhamedov, U., & Webb, S. (2019). Lexical coverage and profiling. *Language Teaching*, *52*(2), 188–200. https://doi.org/10.1017/S0261444819000028

Reynolds, B. L. (2013). Comments on Stuart Webb and J. Macalister's "Is text written for children useful for L2 extensive reading?" *TESOL Quarterly*, *47*(4), 849–852. https://doi.org/10.1002/tesq.145

Ross, S., & Berwick, R. (1991). The acquisition of English affixes through general and specific instructional strategies. *JALT Journal, 13*(2), 131-140. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/jj-13.2-art2.pdf

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33–53.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19*(1), 17–36. https://doi.org/10.1017/S0272263197001022

Sritulanon, A. (2013). The Effects of Morphological Instruction on Reading Abilities of Low Proficiency Adult EFL Learners at a University in Thailand. *LEARN Journal: Language Education and Acquisition Research Network*, *6*(1), 49–65. Retrieved from https://so04.tci-thaijo.org/index.php/LEARN/article/view/102721

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 1–23. https://doi.org/10.1017/S027226312000025X

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System, 37*(3), 461–469. https://doi.org/10.1016/j.system.2009.01.004

（内田　　諭　九州大学）
（森田　光宏　広島市立大学）

「論文」

トピックモデルによる多義性研究：英語動詞 run を例に*

木山　直毅・渋谷　良方

Abstract

Research on polysemy takes a variety of approaches, including studies that are of a more introspective nature grounded in theoretical linguistics (e.g., Pustejovsky, 1995; Lakoff, 1987; Ravin and Leacock, 2002; Tyler & Evans, 2003), as well as those that are more empirically oriented, conducted from the perspectives of corpus linguistics (e.g., Sinclair, 1991; Tognini-Bonelli, 2001) and psycholinguistics (e.g., Kishner & Gibbs, 1996; Gibbs & Matlock, 2001) (for a comprehensive overview of polysemy research, see Nerlich et al., 2003 and Gries, 2015). A common perception in polysemy studies is that the meaning of a word varies with 'context'. The same is true of research in pragmatics. Although context is a useful and important concept, it is also an ambiguous one. This is because there are so many variables involved in it; in addition to co-occurring words, social(-interactional) variables such as the flow of conversation, the context of utterance, and even the age and gender of the speaker can also be part of a context, depending on the theoretical assumption of the study (Stefanowitsch & Gries, 2008, p. 135). The present study reinterprets the notion of context from the viewpoint of topic modeling (e.g., Blei et al., 2003), an NLP method of classifying texts by analyzing the type of topic they are related to, on the basis of the distribution of words. By applying topic modeling, this study explores the polysemy of the English verb *run*. Our chosen topic modeling technique is the one called the Biterm topic model (BTM; e.g., Yan et al., 2013). Based on the analysis of a number of automatically generated topics and the corresponding meanings of *run*, we argue that the context in which the verb is used, or more specifically, the topic in the text, plays a crucial role in identifying the meaning.

1．はじめに

語は一般的に複数の意味を持つ。[1] 例えば，英語の動詞 run を Merriam-Webster（n.d.）で調べると，自動詞用法と他動詞用法を合わせて 30 個に及ぶ意味が見つかる。語が複数の意味を持つ現象は多義性（polysemy）と呼ばれており，この現象についてはこれまで様々な研究が行われてきた。例えば，形式主義の観点からは生成語彙論（Generative Lexicon）による分析（e.g., Pustejovsky, 1995）があり，認知言語学の枠組みからは Lakoff（1987），Ravin and Leacock（2002），Tyler and Evans（2003），Glynn（2016）などがある（認知言語学における，多義性の研究の要約は，Glynn, 2014b 参照）。また，コーパス言語学の分野では Sinclair（1991），Tognini-Bonelli（2001）などがあり，心理言語学においては Gibbs and Matlock（2001），Kishner and Gibbs（1996），Ramsey（2022）などがあげられる（多義性についての包括的な説明は Nerlich et al., 2003 や Gries, 2015 参照）。

上述の通り，語の多義性についてはこれまで様々な方面から多くの研究が行われてきたが，先行研究には共通する認識が見られる。それは，文脈によって意味が変化するという見方である。（なお，文脈によって意味が変化するという考えは多義性研究に特化した認識ではない。例えば, 語用論研究では, 語（あるいは言語表現）の意味を特定するのに文脈が必要であるという認識のもと，意味が理解されるメカニズムが論じられてきた（Bublitz & Norrick, 2011）。）

意味や理解の研究では「文脈」という用語が広く使用されてきたが，文脈という概念について一枚岩的な定義は与えられてこなかった。例えば，文脈は共起語や会話の流れを指すこともあるし，発話状況や，さらには年齢や性といった社会的変数を文脈と呼ぶこともある。すなわち，文脈という用語が指す内容は研究目的によって変わりうるのである（Stefanowitsch & Gries, 2008, p. 135）。

このように,「文脈」とは本来非常に多くの変数（要因）を含む概念である。意味の研究において文脈に説明を求めるアプローチでは，文脈という概念により，具体的にどの要因を考察対象としているのかを明確にするべきであろう。さらに，実証的方法で多義性を扱いたいなら，客観的に規定可能な方法に基づいて文脈を定義すべきである。このような状況を背景として，本研究では，語が用いられる内容，より具体的には，トピックモデルにおけるトピックの観点から（e.g., Blei et al., 2003）文脈を定量化し，語の多義性を記述する方法を提案し，その有用性を論じる。

本稿の構成は次のとおりである。まず第2節では，本稿が想定する理論的背景と，多義性に関する先行研究を概観する。続く第3節では，本稿のリサーチデザインを紹介する。第4節では2つのコーパスを用いた事例研究を紹介し，トピックが英語の動詞の意味とどのような関わりにあるのかを論じる。第5節では，第4節で得られた結果が言語研究にとってどのような理論的含意を持つのかを論じる。第5節ではまた，本稿が使用する手法と先行研究の手法の比較も行う。

2. 背景と先行研究

2.1. 百科事典的意味論

伝統的な意味論（e.g., 形式意味論）では，意味は言語表現と参照物の間の対応，すなわち外延によって規定可能であるという想定の下に研究が行われている（cf., Cann, 1993）。このアプローチとは対照的な意味観を持つのが認知言語学に見られる百科事典的知識に基づくアプローチである。このアプローチでは，人が持つ世界に関わる知識との関連で，語の意味が捉えられている（cf., Haiman, 1980; Fillmore, 1982; 1985）。ここでは一例として，Taylor（2002）による photograph という英語名詞の説明を紹介する。Taylor は，photograph には少なくとも下記の4つの背景知識（ドメイン）が関わると述べている（Taylor, 2002, p. 442）。

（i）写真は，情景を描写したものである。

（ii）写真は，特殊な装置や工程，撮影者の腕前といった様々な技術を通して制作される。

（iii）写真は，通常，紙に現像された映像であるが，映像は硝子板や電子データのような媒体としても存在する。さらに，同じ映像を他の媒体へと複製することが可能である。

（iv）写真は，社会的慣習と密接に結びついたものである。ある人は人生の記念において写真を撮り，ある人は写真で生計を立て，ある人は写真を身分証として使用する。

（1）の事例は，Taylor が上記（i）〜（iv）のドメインが関わる事例としてあげたものである（事例は Taylor, 2002, p. 442 から）。

（1）a. The photograph is torn.
　　b. This is a photograph of me at age 10.
　　c. This photograph has been re-touched.
　　d. The photograph was awarded a prize.
　　e. I'll send you the photograph as an electronic attachment.

Taylor の説明は以下の通りである（pp. 442f）。（1a）では，（i）の知識が前景化される一方で，（ii）と（iv）の知識は背景化されている。（1b）では個人の幼少期が描写されているが，ここでは（iv）の社会文化的慣習が前景化されている。（1c）では修正（re-touch）を可能にするテクノロジー，すなわち（ii）の前半部に関する知識が前景化されている。対照的に，（1d）では撮影者の感性や腕前が関わっており，（ii）の後半部分が前景化されている。（1e）では紙媒体以外の方法で保存できるという（iii）の側面が前景化されている。

上述の 4 つの知識のどれが photograph の意味と関係する背景知識なのだろうか。伝統的な意味論では，いわゆる語用論的意味や知識と語の意味は厳密に区別されるため，photograph の中心的な知識であると考えられる（i）以外は photograph の意味とは無関係だとみなされるかもしれない。一方，百科事典的意味論では，世界に関する知識そのものが語の意味の一部だと考えられることになる（Croft, 1993）。すなわち，百科事典的意味論では，4 つの photograph の理解に関わる背景知識のうち，どれがこの語の言語内的意味で，どれが言語外的意味であるのかを明確に区別することができないと考え，言語に関わる世界の知識が語の意味であると考える（Taylor, 2012）。

2.2. コーパスを用いた多義語の先行研究と本稿の仮説

百科事典的意味観に基づく研究は主に認知言語学の枠組みに依拠する形で行われており，そこではコーパスを通じて実際の言語使用を調べる研究が積極的になされてきた（e.g., Akita, 2012; Glynn, 2009; 2016; Glynn & Robinson, 2014; Heylen et al., 2015; Hoffmann et al., 2019; Lederer, 2019; Schönefeld, 2013; Stefanowitch & Gries, 2003）。それらの中で，本稿に特に関連する分析が Gries（2006）や Divjak and Gries（2006）で提案された behavioral profiles（以下，BP 分析）である。BP 分析では，調査対象となるターゲット語が現れる統語環境に加え，それに伴う形態素，文法的数，アスペクト，意味カテゴリといった極めて幅広い特徴のコーディングがコーパスから収集されたデータに対してなさ

れ，階層的クラスター分析や多次元尺度法といった多変量解析を通じてターゲット語に関する様々な特徴が捉えられる。

BP 分析では，これまで get や run などの高度に多義的な英語動詞の意味記述が行われてきた（e.g., Berez & Gries, 2008; Divjak & Gries, 2008; Glynn, 2014a; Gries, 2006; 2010; Gries & Divjak, 2009; Gries & Otani, 2010; Jansegers & Gries, 2017）。BP 分析の結果は，語の意味を説明するためには語が現れる様々な環境を考慮に入れることが極めて重要であることを明らかにするものであった。上述のように，BP 分析では，数多くの要因についてコーディングがなされる。しかし，同分析法ではコーディングの対象外とされる要因もある。その一つが本研究が注目するトピックである。

本稿が論じる「トピック」とは，語の分布で表現される文書の話題を指す。本研究の仮説は，人が多義語を理解するにあたって，テクスト中で話題になっていること，すなわちトピックの観点から理解しているというものである。この仮説は次のような直感的理由に基づいている。たとえば，陸上の大会という特定のトピックに関する新聞記事の中で英語の動詞 run が用いられる場合について考えてみよう。そこでは，run は「人の高速移動（足をはやく動かすことによる前進移動）（fast pedestrian motion）」の意味で使用される可能性が，他の意味で使用される可能性よりも高いことが予想されよう。一方，選挙に関するトピックで run が現れた場合はどうであろうか。おそらく，その場合には，「選挙での候補者に名を連ねる（to become a candidate）」という意味で用いられている可能性が高いことが予想されるであろう。本稿ではこのような考えに基づき，動詞の意味は語が用いられるトピックとの関係で決まるという仮説を立てている。[2]

では，動詞の意味と，語が用いられるトピックの関係は，どのようなものだろうか。2.1 節で論じたように，百科事典的意味論では，語の意味は，背景知識（ドメイン）の中で指定されると考えられている。たとえば，run の「人の高速移動」の意味を理解する際に，人は関連する背景知識に基づき，その意味を特定する。また，上で挙げた例において，陸上の大会に関する新聞記事の中では，run が「人の高速移動」の意味として使用されやすいであろうという直感は，陸上大会に関する知識が，意味の特定を促進するためのドメインとして働いているからであろう。このように，動詞の意味と，語が用いられるトピックの関係は，動詞の意味とドメインの関係と（ほぼ）同じものとみなすことができるだろう。

上で述べた仮説は，一見すると自明のように思われる。実際，多義性の研究でトピックを暗黙裡に想定しているように解釈可能な研究もある。たとえば，Taylor（2003）は百科事典的意味論と分布意味論（e.g., Turney & Pantel, 2010）の親和性を論じるに当たり，名詞 pen における多義性，すなわち筆記用具（writing implements）と囲い（enclosures）の意味を例にとり，前者ならば直感的に ink，write，paper が共起しやすく，後者ならば children や動物の名前などが共起しやすいだろうと述べている。[3] Taylor の説明を上述の仮説の観点から言い換えると，筆記用具の意味は文房具に関するトピックで現れやすく，囲いの意味は家畜や懲罰などのトピックで現れる可能性が高いと考えることができる。

このように，Taylor（2003）などの先行研究はトピックの観点から解釈し直すことが可能である。しかし，ここで強調したいのは，これらの先行研究では明示的な形でトピックと多義性の関係を論じることはなかったという点である。すなわち，Taylor（2003）のような先行研究においては，語の多義性は主観的で曖昧に定義された（トピックとも解釈可能な）文脈との関係に基づき説明されていたのである。

再現性が担保された形でトピックを論じるには，厳密なアルゴリズムに基づいてトピックを分類する必要がある。今日，コンピュータ技術の発展により，文書のトピックを確率的に計算する手法が考案されている。本稿では次節で紹介するトピックモデルと呼ばれる手法を援用することで上述の仮説の妥当性を検証及び論じていく。

3. リサーチデザイン

3.1. トピックモデル

トピックモデルとは，ある文書がどのようなトピックなのかを解析する手法である。たとえば，次の例が何に関する文書かを考えてみよう。

(2)　従来はペンと紙を持って椅子に座って考える研究者が多かった。しかし，近年はコンピュータの進歩，手法の発展により，膨大な資料の中から実際に使用される言葉の統計学的・確率論的な傾向を見出していくという研究スタイルが重要視されるようになった。

言語研究の歴史を知る者であれば，おそらく，(2）は（コーパス）言語学に関する文書だと理解することができるだろう。しかし，本文にはトピックを表す「言語学」という語は含まれていない。(2）を（コーパス）言語学に関する文書だと理解する者は，本文中の「言葉」や「研究」といった語から内容を判断していると考えられる。トピックモデルとは，語の分布や集合から文書がどのようなトピックと関係するものであるのかを統計的に解析し文書を分類する手法である。

近年のトピックモデルでは，Blei et al.（2003）が提案した latent Dirichlet allocation（LDA）を用いるのが一般的である（Mṻller & Guido, 2016, p. 348)。しかし，LDA は入力する文書データが短い場合には十分にトピックを学習することができず，結果としてトピック抽出の精度が下がるという問題が指摘されている（Tang et al., 2014)。3.3 節にて説明するが，本研究で使用する入力文書はストップワードを含めても 20 語にしかならないため，LDA での解析には不向きである。この問題を解消するために本研究では, 短い入力データのトピック解析に適した biterm topic model（Yan et al., 2013，以下，BTM）を利用する。

筆者らの知る限り，トピックモデルを語彙意味論研究に用いた先行研究はない。しかし, コンコーダンスラインをトピックモデルで解析することによって，ターゲット語がどのようなトピックで用いられているのかを調査することは理論的に可能である。本稿では，トピックモデルによって得られた結果をもとにコンコーダンスラインを精査することで，語のどの意味がどのトピックに現れるのかを調査することができることを示したい。なお, 先に述べたBP分析では，分析者が様々な ID タグを手作業によって付与していく必要があるが，本研究が用いるトピックモデルでは分析の部分はほぼ全て自動化されており，分析の客観性がより高い。

3.2. 使用するコーパス

本稿では 2 つのケーススタディを別々のコーパスを用いて行う。一つ目のケーススタディでは News on Web Corpus（Davies, 2016-, 以下, NOW コーパス）を使用する。本コーパスには 2010 年以降に公開された世界 20 カ国の英字オンラインニュース記事が収集されていて，2022 年 10 月 16 日の時点で約 160 億語が収録されている。すべてのデータを利用することは不可能なため，本研究ではフルテキスト版 2013 年米国データ（約 8,500 万語）に限定し調査を行う。

二つ目のケーススタディでは Corpus of Contemporary American English

（Davies, 2008-，以下，COCA）を使用する。NOW コーパスと同様，COCA の全データを扱うことは不可能なため，ここでは，1990 年から 2017 年までに出版されたフルテキスト版の Fiction データを利用する（約 1 億 1000 万語を収録）。ケーススタディ 1 とは異なるデータソースを用いることで，NOW コーパスとは別な角度からの分析が可能となることが期待できる。

3.3. 手順

本稿では，トピックと語の意味の関連を考察するにあたり，以下の手順で調査を行った。まず，共起語の取得範囲は，前後 10 語，合計で 20 語を抽出した。また，いくら BTM が短い文書解析に優れた手法とはいえ（3.1 節参照），1 語や 2 語から構成される文書の解析は極めて困難なため，ストップワードを除外した後に 5 語未満となったデータを除外した。[4]

本研究ではデータ収集については，従来のコーパスを用いた多義性の研究とは異なった収集方法を採用した。コーパスを用いた意味研究では，文中に共起する語を調査するのが一般的である。たとえば動詞の意味研究の場合は主語や目的語，あるいはその修飾語などに注目することから，共起語の検索は対象となる文を超えることはあまりない。しかし，本研究は語が現れるトピックを扱うことから，文を超えてデータ収集を行っている点に注意されたい。

以上の手続きに基づき，本研究では英語の動詞 run をケーススタディとして扱う。本稿の冒頭でも述べたように，英語の動詞 run は辞書に 30 以上もの意味が収録される語であり，これまでも百科事典的意味論の立場から様々な調査がなされてきた（e.g., Gries, 2006; Langacker, 1988）。本稿では，これらの先行研究との比較のために run を調査対象とした。

4.　分析結果

4.1. NOW コーパスにおける run

4.1.1. 最適トピック数

トピックモデルでは，トピックの分類は自動的に行われるものの，トピックの数については分析者が決定する必要がある。最適なトピック数を決めるにあたり，LDA では，トピック数の妥当性を示す診断（diagnosis）等を利用する場合がある。しかし，BTM は文書の生成プロセスを計算しない手法であり，これらの指標を使用することができない（Yan et al., 2013, p. 1449）。ゆえに，

本研究では，排他性（exclusivity）を利用することにした。

排他性とは，トピックモデルで計算したトピックの質を表す指標である。この指標は，トピックごとの単語の出現率（ø 値）を，トピック全体の ø 値の和で割った値の平均値である。排他性は，値が高いほど，各トピックの特徴を表していると評価される。図 1 は, トピック数を 5 から 50 の範囲で 5 毎に計算し，各トピックにおいて，高出現率語 10 語を用いて排他性を算出した結果である。この図から，トピック数を 15 と 40 に設定したときに高い値を示していることがわかる。ここでは, これらのトピックを質的に解釈し, 比較した結果, トピック数を 15 に設定するほうが解釈しやすい結果であったことから，排他性が最も高い値を採用した。

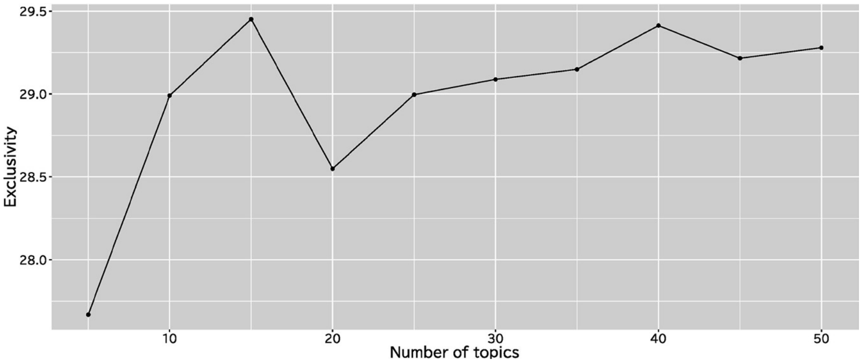


図 1. Topic 5 から 50 の排他性（上位 10 語を使用）

なお，BTM では，計算の過程でテキスト中に幅広く使用される語を Topic 1 に集めることで，残りのトピックの精度が高められる。そのため，Topic 1 は研究においては利用されないトピックであり（background），本稿でも Topic 1 は分析対象からは除外し，14 のトピックで調査を行った。

4.1.2. BTM の結果

3.3 節で述べた手続きを通じて得たデータに対し，BTM で 14 のトピックを算出した結果を表 1 に示す（紙幅の都合により，一部抜粋）。この表の左端には各トピック番号を記載している。左から 2 列目には，排他性の算出時に用いた語彙を ø 値の降順に並べ，筆者らが解釈した各トピック名を 3 列目，そして各トピックで使用される run の意味については一番右の列に記述している。な

お，本研究では（フルテキスト版）コーパスに付与されたPOSタグを含めた調査を行っているが，本稿では視認性を高めるためそれらを除外して表示している。

表 1. トピックモデルによる分析結果

	トピック構成語上位 10 語...	トピック名	run の意味
Topic 2	power, gas, energy, fuel, natural, battery, plants, electricity, batteries, water	エネルギー	To execute/function
Topic 3	people, run, time, really, marathon, day, running, good, things, way	マラソン	Fast pedestrian motion
Topic 5	police, away, man, house, away, ran, toward, people, saw, old	犯罪	To escape
Topic 13	money, government, percent, fund, million, billion, social, dollars, security, deficit	政府と資金	To become used up
Topic 14	water, blood, run, face, hair, hot, hands, air, bar, room	液体の流れ	To flow
Topic 15	line, water, river, north, road, city, east, miles, aground, south	地理的概念	To flow (fictive motion)

次に，実際に各トピックにおいてrunがどのように使用されているのかを確認していく。[5] 表1とはトピックの順番が前後するが，文字通りの移動（literal motion）を表すトピックから見ていく。まずTopic 3はマラソンに関するトピックだと思われる。Topic 3で用いられる語と共起する事例を見てみると，マラソントピックで用いられる場合，runは「人の高速移動」である傾向が強い。

(3) a. ... if you have not been exercising, don’t expect to ***run*** a half marathon with only a couple weeks of running.
  b. The Boston Marathon is America’s iconic race, the oldest marathon in the country, and the most important. Eighteen people ***ran*** it in 1897; last year, thirty-five thousand did.

次のTopic 5は，犯罪に関わるトピックだとみなせる。このようなトピックでは，次のような「逃走（to escape）」の意味で用いられる例が見られる。

(4) a. ··· 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.

b. Before Janelle turned 15, she was suspended from school, thrown out of her house by her parents, ***ran*** away from her foster home, and got arrested for shoplifting.

（4a）は，カーテンを開けた際に逃亡中の凶悪犯が警察に追われている状況を見たことが表されている。また（4b）では，Janelle が成長過程において様々な犯罪に手を染めてきたこと，そして彼女が里親から逃げた事象が述べられている。これらの事例で見られる用法は，先に見た「人の高速移動」と類似したものであり，文字通りの移動を表す点では同一の意味だと考えることができる。一方で，スポーツ，特に長距離走における移動では，誰から逃げるのか（追跡者）の存在が明示化されないが，（4）では警察や里親といった追跡者の存在が何らかの形で表される。このように（3）と（4）とでは，事態に関わる参与者が異なっていることがわかる。

次に Topic 2 では，natural gas や water power といった語が現れていることからエネルギーに関わるトピックであることがわかる。このトピックでの run は，Topic 3 や Topic 5 で見た文字通りの移動とは異なり，（5）の事例が示すように，「機械が作動する（to function）」という抽象的な意味で用いられている。

（5）a. The devices ***run*** on natural gas and produce electricity with fewer emissions than a diesel gen-set.

b. 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 …

Topic 14 と Topic 15 は，いずれも液体に関連するトピックであり，一見するといずれも「液体の移動（movement of liquid）」の意味のように見えるが，実際の用例を見てみると興味深い差が見つかる。（6）は，Topic 14 の事例であるが，いずれの例も液体が移動する経路を表す前置詞（from, down, over）が現れており，熱湯の移動を表している。このことから，Topic 14 の run は液体の移動事象を表していると言える。

（6）a. Hot water from the pool above ***ran*** down the bank of the terrace, which was striated in several colors...

b. When it’s time to serve, ***run*** hot tap water over the container to release the ice.

一方で Topic 15 の例に現れる語は，地理的な概念であり，これらが物理的に移動することは考えにくい。むしろ, 概念化者が川や下水の経路を心的にたどった仮想移動（fictive motion）（Talmy, 2000, cf., Langacker, 1986; Matsumoto, 1996）の一種だと考えるべきであろう。

（7）a. Labeled “a state scenic river, a Catoctin Creek near Waterford ***runs*** under a single lane bridge on a lightly traveled road ...

b. And yet its combined sewer system was constantly overflowing, in 400 places over 375 square miles, ***running*** afoul of federal water quality standards.

以上，トピックと語の意味の間には関係が成り立つことを示したが，ここで 1 つ疑問が生じる。本研究では 14 個のトピックを立てているため，14 個のドメインを見出したことになるが，これらの背景知識は互いにどのような関係にあるのだろうか。本稿では，語の意味はトピックとの関係で決定されると考えることから（2.2 節参照），ここではトピックの類似度を計算することで複数の意味間の関係を考えていく。図 2 は，確率分布の類似度を測る指標として Jensen-Shannon divergence を用いた結果を多次元尺度構成法（MDS）に基づき描画したものである。

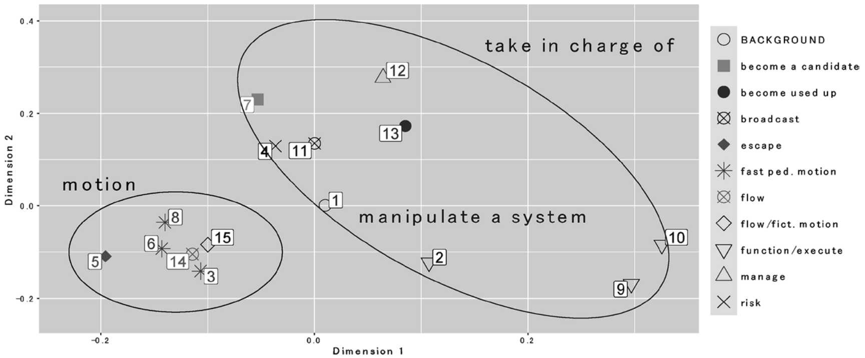


図 2. MDS による run の意味の類似度（1–2 軸）

図 2 では, 第 3 象限において「人の高速移動」(=Topic 3, 6, 8) や「逃走」(=Topic 5）といった文字通りの移動の意味が集まっている。また，Topic 14，15 の「物質の流れ」と「仮想移動」もそれらと同じまとまりに位置している。これらの

分布から，左下（青）の楕円で示す領域は run の「移動（motion）」の意味でまとめることができるだろう。ただし，2 軸と 3 軸（図 3）を見てみると，「物質の流れ」（Topic 14）と「仮想移動」（Topic 15）は第 2 象限に現れており，文字通りの移動の意味と離れたところに位置している。また，「逃走」（Topic 5）が「人の高速移動」（Topic 3，6，8）と離れて位置しているのも確認でき，各意味には一定の差があることが確認できる。図 2 の第 4 象限には，「機械が作動する（to function/execute）」の意味がまとまり，図 2 の Y 軸上の正の位置には「選挙に出馬する」や「経営する（to manage a company/business）」といった抽象的な意味が集まり，何かを「統括する（to take in charge of）」意味が集まっている。

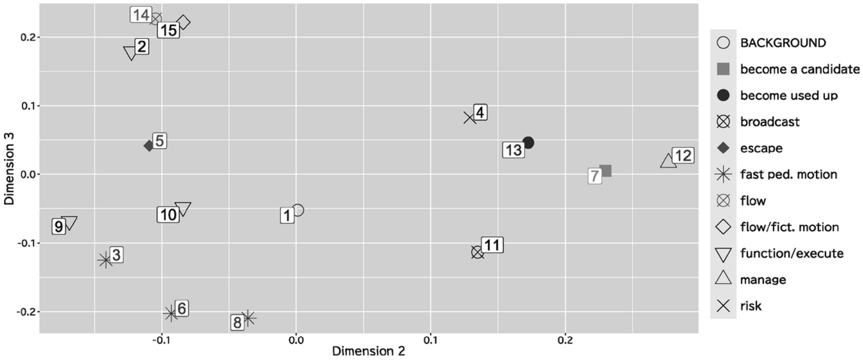


図 3. MDS による run の意味の類似度（2–3 軸）

以上の分布から，NOW コーパスを BTM で解析し，トピックから run の意味を記述すると（i）移動（motion）（図 2 の左下（青）の楕円）と（ii）複雑なモノを機能させる（to cause complex entities to work）（図 2 の右（赤）の楕円）の意味に大別することができる。

4.2. COCA における run

4.2.1. 最適トピック数

Fiction データでは, 4.1.1 節と同じ条件で実験し, トピックの最適値を探った。図 4 は，その結果である。排他性は，トピック数が 50 の場合に最大値となっているが，数理的に理想的なトピック数が必ずしも人間にとって解釈しやすいとは限らない（Chang et al., 2009; Törnberg & Törnberg, 2016）。そのため，本研

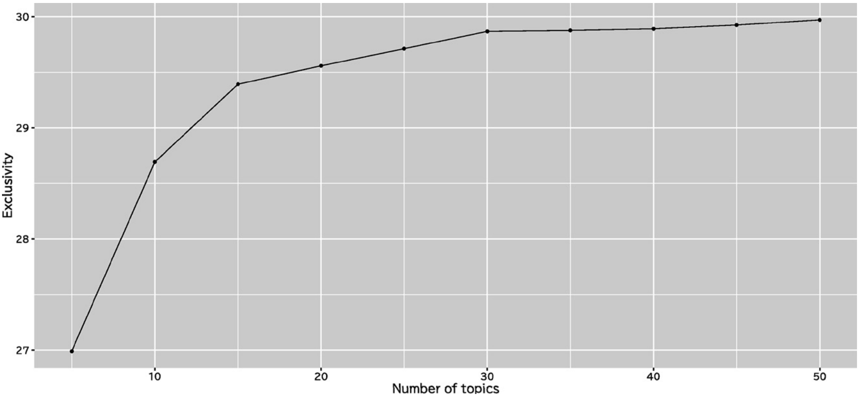


図 4. Topic 5 から 50 の排他性

究では，値の上昇が緩やかになる30から50の間を質的に解釈し，最もトピックの特徴が現れていたトピック数40を最適な数と判断した。

4.2.2. BTM の結果

Fiction データを BTM で調査すると，4.1.2 節では見られなかった身体を表すトピックや建造物に関するトピックが確認できた。以下では，これら2つのトピックの特徴を見ていく。

まずは身体を表す表現について述べる。表2がそのリストである。Topic 11 の構成語を含む事例を見ると，（8）に示すように特定の感覚，あるいは感情へと変化する様子が描かれていることが分かる。そこで，この意味を「感覚・感情の瞬間的変化（an instantaneous change of emotion/feeling）」と呼ぶことにする。寒気や震え，恐れといった感覚・感情は瞬間的に生じるものであり，この瞬間的な感情変化と run が持つ高速移動の様との類似性から，この表現が生じた可能性がある。

（8） a. Erin grimaced in fear. I felt a tingle ***run*** up my spine.

b. He’d always envisioned her as [⋯] pale from fear or from the chill that ***ran*** along her spine.

c. She froze, her body ***running*** cold and clammy.

d. Zayder felt fear ***run*** in harsh prickles across his own skin.

表 2. Fiction における身体部位を含むトピック

	トピック構成語上位 10 語...	トピック名	run の意味
Topic 11	felt, spine, body, shiver, chill, moment, mind, cold, fear, eyes	感情や感覚	An instantaneous change of emotion/feeling
Topic 35	hand, finger, fingers, arm, left, scar, cheek, side, shoulder, neck	接触する身体部位と接触対象	To touch an object by moving a body part
Topic 37	hair, hand, fingers, hand, head, black, thick, brown, face, short	接触する身体部位と接触対象	To touch an object by moving a body part
Topic 39	face, tears, eyes, blood, hands, cheeks, sweat, mouth, nose, tongue	身体部位と液体	Body fluids flow on a body part

e. She closed her lips around my tongue, and a shiver ***ran*** through my body.

Topic 35 は，finger(s) や arm などの動かす対象の身体部位と，cheek や neck といった手で触れる身体部位を表す単語から構成される。そこで，本トピックを「接触する身体部位と接触対象の身体部位」とラベリングする。これらの語を含む事例には，身体部位の移動，特に（9a）のように「何かをなぞる」という意味を語用論的に示唆するものが含まれる。

（9）a. I fetch my backpack, don’t need unzip it cause the zipper don’t work anyway, and pull out my book. It feel heavy and cool in my hand, and I ***run*** my finger cross the shiny silver letter of the title.

b. She leaned in to rest her cheek against his and he ***ran*** a hand down her bare back, relaxing it above the concave of her spine.

c. She palmed her neck. She ***ran*** her hands up and down her forearms, brushing up the gold hairs, then smoothing them again.

Topic 37 は fingers や hand といった身体部位を含む点で Topic 35 に類似している。しかし, 接触対象となるものは, Topic 35 では（表 2 には現れていないが）letter（「文字」の意味）や scar といった身体部位以外の語彙も見られるが，Topic 37 は hair や brown，short といった髪の毛に関連する語や身体部位が多く含まれる。そのため，Topic 37 には「接触する身体部位と接触対象の身体部位」

というトピック名をつけた。これらの構成語が用いられる事例は次のようなものである。

（10）a. She ***ran*** her fingers through her long, dark brown hair ...
　　b. ... Lou self-consciously ***ran*** a hand up and down the sides of his face.

本トピックとTopic 35との間には，fingerの文法的数の違いが確認できる。Topic 35の場合，fingerは単数形・複数形の両方が現れているが，Topic 37では複数形しか現れていない。Topic 35の場合，（9）で見たように，触れる対象が傷であったり文字であったりと，（もちろん，複数の指で触れるものもあるが）一般的に一本の指で触れるような接触対象物が含まれる。一方，Topic 37の場合は髪の毛や頭など，指一本で触れることは稀であるものが含まれている。このように，runの意味をBTMで調査すると，対象物をどのように触れるのかという（いわゆる）言語外知識をも反映する結果を得ることができた。

表2の最後のTopic 39は，handsやcheeksなどの身体部位に加え，tearsやblood, sweatといった体液が含まれている。これらの分布から，Topic 39を「身体部位と体液」であると判断できる。これらの語が使用される事例は次のとおりである。

（11）a. Beads of sweat ringed his sunburnt head in a crown-like way and ***ran*** down into his watery faded-blue eyes.
　　b. You are crying, and your tears silently, uncontrollably ***run*** down your cheeks, along your nose, to your mouth.

（11）の事例が示すように，Topic 39は人間の身体から出てくる液体についてのトピックであり，runは，「体液が身体部位を流れる」という意味で用いられていることが分かる。

以上のように，表2に挙げる全てのトピックにおいて，身体部位を表す表現が出ており，またトピック自体も身体に関わるものであることは間違いないであろう。しかし，各トピックのコンコーダンスラインを観察すると，runの意味は，それぞれで異なっていることが明らかになった。特に，Topic 35と37で確認した（いわゆる）言語外的知識の差に関しては，トピックの差異が語の意味の差を生んでいると言えるだろう。

次に，建造物に関するトピックについて見ていく（表 3）。この表から，Fiction データにおける建造物に関するトピックでは，建造物（e.g., house, street）や内装（e.g., ext/int(erior), door）に関わる語が多く集まっていることがわかる。

表 3. Fiction における建造物に関わるトピック

	トピック構成語上位 10 語...	run の意味
Topic 3	toward, street, back, door, ext, night, man, starts, turns, people	Fast pedestrian motion
Topic 21	door, back, room, house, toward, front, came, stairs, kitchen, back	Fast pedestrian motion
Topic 24	int, night, room, ext, door, house, street, stairs, hallway, phone	Fast pedestrian motion

なお，これらの各トピックに固有のラベルをつけるのは非常に困難に思われる。なぜなら，構成語の分布が高度に類似しているからである。[6] そのため，本稿ではこれらのトピックについては「建造物」という総称的用語を用いて run の意味を調査する。建造物トピックでは，次のような run の事例が見られる。

（12）a. I ***ran*** back through the dining room and toward the front door.

b. As Lisa lay there, shaken from the noise, she heard Aunt Mary call, “Lisa, I need your help!” Lisa ***ran*** down the stairs to the kitchen ⋯

上記の事例が示すように，COCA の Fiction データにおいて，run が建造物トピックで用いられる場合，建造物が移動の始点あるいは着点となっている事例が目立つ。これらの例から，建造物トピックで現れる run の意味は「人の高速移動」であることがわかる。

この「人の高速移動」の意味は，NOW コーパスにおいても run の意味として確認することができたが（4.1.2 節），コーパス間で言語表現として用いられる構文が異なっている。NOW コーパスでは，「人の高速移動」はスポーツの文脈で用いられることが多く，結果として（3）のように出場する競技や，run for 376 yards のように走る距離が後続する事例が多かった。一方，Fiction データだと，（12）に示したように run back や run through，run down A to B のように経路あるいは目的地が後続する事例が多く見つかる。この違いは，語の意味としては「人の高速移動」と分類可能であっても，トピックやレジスターが異なると，移動事象のどの部分に焦点を当てているのかが異なることを示している。[7]

5. 結果の解釈と結論

前節では，run の意味とトピックとの間に関連性があるという分析結果を報告した。より具体的には，BTM で得た NOW コーパスと COCA の Fiction データとの分析結果を比較すると, 共通して現れる run の意味も存在するが, 各コーパスの特徴を表すものも確認された。

まず，コーパス間で異なって見られた意味について考察する。COCA の Fiction データに特徴的な意味は，（9）や（10）に示した「身体部位を動かす」であった。これは，NOW コーパスを用いた本研究では確認されなかった用法であり, 使用されるレジスターに偏りがあることを示唆している。Gries（2006）では，コーパスに含まれる様々なレジスターを統一的に扱っており，意味におけるレジスターの偏りについては言及していない。本研究により，レジスター間の意味の偏りを考慮に入れる必要性が示唆される。

次に，分析結果では，両コーパスに共通して現れる「人の高速移動」の意味においても重要な差が見られた。すでに論じたように，NOW コーパスでは目的語に出場競技をとる用法や，どのくらいの距離を走ったのかを示す例やトピックが目立ったのに対し，COCA の Fiction データでは経路を表す表現が目立った。これは「人の高速移動」という 1 つの事象に対して人が持つ百科事典的知識のうち，異なった側面を前景化していると言える（Fillmore, 1982 参照）。この点も，語の使用がレジスターに偏りがあることを示唆している。

以上の点は，2 つのレジスターで言葉の使い方が明らかに異なっており，また使用される語の意味も異なっていることを示している。このことは，用法基盤モデルや百科事典的意味論においては，意味とレジスター間の深い関係性を論じる必要があることを強く示唆するものであり，言語の意味を社会・文化的観点から規定する必要性を主張する認知言語学の社会的転回（social turn; Harder, 2010, p. 3）の方向性を支持するものである。

次に，方法論の点から，本研究の研究結果について述べておく。特に，BTM と Gries らが提案する BP 分析との違いに着目して論じる。まず，BTM を用いることのメリットの 1 つに，分析に要する所要時間の短さと客観的信頼性の高さがある。BP 分析では，例えば，動詞の意味を分析する場合には，相や時制に関する形態素タグ, 動詞が生じる節や句の種類や文のタイプ（平叙文, 命令文など）といった統語タグ，動詞の共起語の意味的特徴（e.g., 主語や目的語が生物かどうか），そしてターゲット語の意味などの情報が分析者によって

付与されていく。Gries（2006）では，815 の run の例に対して，252 個の ID タグが付与されているが，この内の 40% が手作業による形態，統語，意味のタグである。この作業は膨大な時間，労力がかかる上，タグの信頼性の確保が難しい。しかし，BTM であれば，これらの作業は不要である。これは，意味分類を手早く，また一定の客観的信頼性を保ちながら行うことを目指す分析者にとっては利点であろう（ただし，両者は競合する手法というわけではなく，補い合うものであると考えられる）。

2 つ目に，トピックというこれまでの多義性研究ではあまり考慮されてこなかった変数を意味分析の中心に位置付けることは，意味とトピックの関係を新たに探究する上で意義深いものと思われる。本研究が論じたように，意味とトピックには不可分な関係があり，曖昧な概念となりがちな文脈をトピックの観点から捉え直すことにより，新しい意味研究を切り拓くきっかけとなるかもしれない。

3 つ目に，コロケーション統計との比較から，BTM の利点を述べる。T 検定やカイ二乗検定量を用いたコロケーション分析では，例えば hair と fingers が run にとって重要なコロケーションだったとしても，run と hair，run と fingers のように 2 語の関係しか捉えることができず，3 語以上の関係を見ることはできない。しかし，BTM はトピックでの分析を行うため，hair と fingers が run とどのような背景知識を喚起するのかを分析することが可能である（4.2.2 節参照）。BTM が持つこの特徴は，言語に関わる世界の知識を統計的に浮かび上がらせる手法として有用だと言えよう。

なお，多くの研究と同様，BTM にも当然デメリットもある。まず，BTM や LDA は，実験のたびに結果が少しずつ変化するが，BP 分析は常に一定である（ただし，BP 分析でも，特に意味のタグ付けに関しては分析者が異なれば結果は一定とはならない）。また，BTM は，文書（あるいはコンコーダンスライン）のトピック推定精度が低く，各トピックを構成する語を使って用例を探していく必要がある（注 5 参照）。一方，BP 分析では，すべてが手作業でタグ付与されるので，関連するデータを量的に調査することができる点については BP 分析に軍配が上がる。

本研究の結論は，以下のとおりである。表 1 ～表 3 に示したように，英語の動詞 run が使われるトピックを観察することができ，そしてそのトピックと動詞の意味との間に関連を見出すことができた。また，レジスター別のコーパスを利用し，動詞 run の用法を比較することで，語が前景化する百科事典的知識

の差も明らかにすることができた。BTMを用いることで, 語の意味研究に加え, 語の用法とレジスターの関係も分析することが可能となる。もちろん，BTMはまだ新しい技術であり，コンコーダンスラインの予測精度が低いという弱点はあるが，意味研究への新たな視点を導入することができる点で，有用な手法であると言えるだろう。

注

* 本稿は，『英語コーパス研究』の3名の査読者より，有益なコメントをいただいた。記してお礼申し上げる。なお，本研究は，英語コーパス学会第47回大会での発表内容を発展させたものである。また，本研究は，科研費（基盤C研究課題20K00667）の助成を受けている。

1 本稿では，意義（sense）ではなく，意味（meaning）という用語を用いる。

2 すべての語の意味がトピックと関連するとは想定できず，前置詞などの文法化の進んだ語はトピックとの関連は薄いと思われる。この検証は今後の課題としたい。

3 Taylor（2003）の例は，同音異義語とも考えられるが，同音異義語と多義性との差は連続的であると考えられる（国広，1982）。

4 トピックモデルでは，ストップワードを除外するべきではないとする立場もある（Brookes & McEnery, 2019）。しかし，ストップワードを含めた場合，筆者らが行った実験では，多くのトピックが機能語で構成され，解釈することができなかった。そのため，本稿では，ストップワードを除外した。

5 BTMでは，トピック学習と文書トピックの推定は，別の手順で計算がなされる。具体的には，(i) コンコーダンスライン全体を入力データとしてトピックを学習し，(ii)（i）で学習したトピックを手がかりに，各コンコーダンスラインのトピックを推定する。表1や表2で示す結果は（i）の結果であるが，(ii) の推測精度が低いため，(i) で得られた結果をもとに，分析者がコンコーダンスラインを探していく必要がある。

6 査読者より，各トピックに固有のラベルをつけるのが困難になった理由として，設定するトピック数が多すぎることが原因であると考えられ，「トピックの最適数を決める手法に問題はないのか」という指摘を受けた。LDAでは，複数の指標を組み合わせて，総合的に判断するのが一般的であるが，BTMは，計算プロセスの特性から，最適なトピック数を決める数理的な手法は，あまり多く提案されていない（4.1.1節）。しかし，排他性は，BTMだけでなく，LDAなどでも広く使用されるため，一定の信頼性はあるものと考えられる。

7 コーパス言語学の伝統では,「テクストジャンル」と「レジスター」は，入れ替え可能な用語として用いられてきた。一方で，それぞれの用語に対して，一貫した定義が与えられてこなかった（Biber & Conrad, 2009; Conrad, 2015; Seoane & Biber, 2021）。そのため，本稿では，総称的な用語として「レジスター」を用いる。

参考文献

Akita, K. (2012). Toward a frame-semantic definition of sound-symbolic words: A collocational analysis of Japanese mimetics. *Cognitive Linguistics*, *23*(1), 67–90.

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 NWLC*, 157–166.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

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

Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, *21*(3), 3–21.

Bublitz, W., & Norrick, N. R. (2011). Introduction: The burgeoning field of pragmatics. In W. Bublitz & N. R. Norrick (Eds.), *Foundations of pragmatics* (pp. 1–20). Mouton de Gruyter.

Cann, R. (1993). *Formal semantics: An introduction*. Cambridge: Cambridge University Press.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09)*, 288–296.

Conrad, S. (2015). Register variation. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 309–329). Cambridge University Press.

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

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

Davies, M. (2016-). *Corpus of News on the Web (NOW)*. Available online at https://corpus.byu.edu/now/.

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

Divjak, D., & Gries, S. (2008). Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon*, *3*(2), 188–213.

Fillmore, C. J. (1982). Frame semantics. In The Linguistics Society of Korea (Ed.), *Linguistics in the morning calm* (pp. 111–137). Hanshin Publishing Company.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, *6*(2), 222–254.

Gibbs, R. W., & Matlock, T. (2001). Psycholinguistic perspectives on polysemy. In H. Cuyckens & B. E. Zawada (Eds.), *Polysemy in cognitive linguistics: Selected papers from the international cognitive linguistics conference, Amsterdam, 1997* (pp. 213–239). John Benjamins.

Glynn, D. (2009). Polysemy, syntax, and variation a usage-based method for cognitive semantics. In V. Evans & S. Pourcel (Eds.), *New directions in cognitive linguistics* (pp.

77–104). John Benjamins.

Glynn, D. (2014a). The many uses of *run*: Corpus methods and socio-cognitive semantics. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 117–144). John Benjamins.

Glynn, D. (2014b). Polysemy and synonymy: Cognitive theory and corpus method. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 7–38). John Benjamins.

Glynn, D. (2016). Quantifying polysemy: Corpus methodology for prototype theory. *Folia Linguistica, 50*(2), 413–447.

Glynn, D., & Robinson, J. A. (Eds.). (2014). *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*. John Benjamins.

Gries, S. T. (2006). Corpus-based methods and cognitive semantics: The many senses of *to run*. In S. T. 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: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon, 5*(3), 323–346.

Gries, S. T. (2015). Polysemy. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 472–490). Mouton de Gruyter.

Gries, S. T., & Divjak, D. (2009). Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In V. Evans & S. Pourcel (Eds.), *New directions in cognitive linguistics* (pp. 57–75). John Benjamins.

Gries, S. T., & Otani, N. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, *34*, 121–150.

Haiman, J. (1980). Dictionaries and encyclopedias. *Lingua*, *50*, 329–357.

Harder, P. (2010). *Meaning in mind and society: A functional contribution to the social turn in cognitive linguistics*. Mouton de Gruyter.

Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, *157*, 153–172.

Hoffmann, T., Horsch, J., & Brunner, T. (2019). The more data, the better: A usage-based account of the English comparative correlative construction. *Cognitive Linguistics*, *30*(1), 1–36.

Jansegers, M., & Gries, S. T. (2017). Towards a dynamic behavioral profile: A diachronic study of polysemous *Sentir* in Spanish. *Corpus Linguistics and Linguistic Theory*, *16*(1), 145–187.

Kishner, J. M., & Gibbs, R. W. (1996). How "just" gets its meanings: Polysemy and context in psychological semantics. *Language and Speech*, *39*(1), 19–36.

国広哲弥．（1982）．『意味論の方法』．大修館書店．

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Langacker, R. W. (1986). Abstract motion. In *Berkeley linguistics society (BLS)* (Vol. 12, pp. 455–471).

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

Lederer, J. (2019). Lexico-grammatical alignment in metaphor construal. *Cognitive Linguistics*, *30*(1), 165–203.

Matsumoto, Y. (1996). Subjective motion and English and Japanese verbs. *Cognitive Linguistics*, *7*(2), 183–226.

Merriam-Webster. (n.d.). Run. Retrieved September 23, 2022, from https://www.merriam-webster.com/dictionary/run

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python*. O'Reilly.

Nerlich, B., Todd, Z., Herman, V., & Clarke, D. D. (Eds.). (2003). *Polysemy: Flexible patterns of meaning in mind and language*. Mouton de Gruyter.

Pustejovsky, J. (1995). *The generative lexicon*. The MIT Press.

Ramsey, R. E. (2022). Individual differences in word senses. *Cognitive Linguistics*, *33*(1), 65–93.

Ravin, Y., & Leacock, C. (Eds.). (2002). *Polysemy: Theoretical and computational approaches*. Oxford University Press.

Schönefeld, D. (2013). *It is ... quite common for theoretical predictions to go untested* (BNC_CMH). A register-specific analysis of the English *go un-V-en* construction. *Journal of Pragmatics*, *52*, 17–33.

Seoane, E., & Biber, D. (2021). A corpus-based approach to register variation. In E. Seoane & D. Biber (Eds.), *Corpus-based approaches to register variation* (pp. 1–17). John Benjamins.

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

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.

Stefanowitsch, A., & Gries, S. T. (2008). Channel and constructional meaning: A collostructional case study. In G. Kristiansen & R. Dirven (Eds.), *Cognitive sociolinguistics: Language variation, cultural models, social systems* (pp. 129–152). Mouton de Gruyter.

Talmy, L. (2000). *Toward a cognitive semantics, volume 1: Concept structuring systems*. The MIT Press.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st international conference on machine learning* (pp. 190–198).

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

Taylor, J. R. (2003). Polysemy's paradoxes. *Language Science*, *25*(6), 637–655.

Taylor, J. R. (2012). Contextual salience, domains, and active zones. In H.-J. Schmid (Ed.), *Cognitive pragmatics* (pp. 151–174). Mouton de Gruyter.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Törnberg, A., & Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society*, *27*(4), 1–22.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Tyler, A., & Evans, V. (2003). *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on world wide web* (pp. 1445–1456).

（木山　直毅　北九州市立大学）

（渋谷　良方　金沢大学）

「論文」

On the *NP as we know it* Expression and its Variants

Yoshiaki SATO

Abstract

This paper quantitively and qualitatively investigates the overall distribution and behavior of the *NP as we know it* expression (e.g., *the world as we know it*) and its variants (e.g., *records as we knew them* and *his life as he knows it*), by employing the Corpus of Contemporary American English (COCA). It reveals that this group of expressions significantly tends to be used concerning the following three situations: "origination," "transformation," and "non-existence." Examples relating to the "non-existence" of the referent of a nominal modified by an *as*-clause whose verb is *know* (e.g., *The world as we know it will be over*) occur by far the most frequently; these account for half of the total. Moreover, it is claimed that there are two functional restrictions on the use of this kind of nominal: (a) the referent of the nominal in question is very unlikely to cause another entity to change its state and (b) the attributive or identificational information predicated of this referent needs to be unexpected or unfamiliar to hearers or readers.

1. Introduction

There is a nominal expression with an *as*-clause whose verb is specified as *know*, as shown in (1).

(1) The world as we know it will cease to exist.
(COCA, Fiction, 2007; emphasis mine; the same applies hereafter)

In (1), the *as*-clause *as we know it* modifies the nominal *the world*, and the resulting more complex nominal *the world as we know it* functions as the subject of the sentence, which depicts the referent of this nominal as being lost in the future. Despite interesting syntactic and semantic properties of this kind of nominal (cf. Section 2), a detailed investigation to reveal its effects and restrictions has hardly been conducted. To clarify these points, this paper focuses on the nominal's collocational and functional aspects. It shows that the referent of a nominal modified by an *as*-clause whose verb is *know* significantly tends to participate in situations related to its "origination," "transformation," and (especially) "non-existence." It also points out two functional restrictions on the nominal: the *as*-clause tends to severely avoid modifying a nominal whose referent causes a change in another entity's state, and the attributive or identificational information predicated of the referent of the nominal in question needs to be unexpected or unfamiliar to hearers or readers.

This paper is organized as follows: Section 2 highlights several observations on the nominal modifier *as*; Section 3 conducts a corpus-based investigation of a nominal modified by an *as*-clause whose verb is specified as *know*; Section 4 attempts to account for its results and points out two functional restrictions on this nominal; and Section 5 concludes the study and offers suggestions for further research.

2. General Characteristics of the Nominal Modifier *As*

This section explores the characteristics of the nominal modifier *as*, mainly based on Yagi (1996), who summarizes the findings of previous studies and discusses the functions of *as* (see also Kanaguchi (1978), Kinugasa (1979), Fukumura (1985a, b), Ogawa (1985), Hirota (1988)).

To begin with, let us look at the pronoun included in the *as*-clause. While the pronoun co-referential with the preceding nominal must be overtly expressed in the *as*-clause, this is not the case in its paraphrase using the relative pronoun *that*, as seen in (2).

(2) a. The world$_i$ as/*that we know it$_i$ will cease to exist.
b. The world *as/that we know will cease to exist.

Moreover, the pronoun *it* in (3a) functions as the object of the verb *know*, but it does not refer to the propositional content of the matrix clause (cf. Huddleston and Pullum (2002)) as opposed to (3b), which can be interpreted as "we know that language is a unique human property."

(3) a. Language as we know it is a unique human property.
b. Language, as we know, is a unique human property. (Yagi (1996: 213))

Yagi (1996) argues that due to its syntactic behavior, *as* in this type of expression is a nominal modifier. The *as*-clause and the preceding nominal form a constituent that can occur within the focus position in the *it-cleft* construction, as seen in (4).

(4) a. The novel as I have described it has never been established in America.
(L. Trilling, *The Liberal Imagination*) (Kanaguchi 1978: 89)
b. It is the novel as I have described it that has never been established in America.
(Yagi (1996: 217))

Notably, any word string cannot intervene between the *as*-clause and the preceding nominal without breaking this constituency. The *as*-clause in (5a) does not modify the preceding nominal (i.e., *the novel*), but it functions as an adverbial adjunct. In (5b), the *as*-clause is interpreted to modify *a unique human property*, not *language.*

(5) a. You are describing the novel exactly as I have described it.
b. ?Language is a unique human property as we know it. (Yagi (1996: 212, 217))

Therefore, Yagi (1996) (and the other previous studies) deal(s) with the relationship between the *as*-clause and the preceding nominal, but not the one between the complex nominal modified by this *as*-clause and the matrix clause.

With respect to this point, the descriptions of the *NP as we know it* expression in some dictionaries provide beneficial information. Specifically, the descriptions in the 9th edition of *Collins COBUILD Advanced Learner's Dictionary* (CCALD9) and the 2nd edition of *Macmillan English Dictionary for Advanced Learners* (MEDAL2) are worth noting, as shown in (6).

(6) a. the form of a thing or system in which it exists now and which is familiar to most people (CCALD9)
b. something that people are familiar with, especially something that is likely

to change (MEDAL2)

These descriptions are considered to supplement each other. In (6a), there are two important elements: *now* and *most people*. As will be discussed in Section 4, the current form of a certain entity is at issue and the pronoun *we* is likely to be interpreted as referring to people in general rather than particular persons, including the speaker. While (6a) does not clarify any collocational preference of the phrase, (6b) overtly states that the referent of a nominal modified by the *as*-clause is likely to change. Regrettably, it does not identify what type of change is most likely to happen.[1] Furthermore, the overall distribution and behavior of this kind of expression are far from clear and remain to be elucidated. Hence the need for such a survey, which will be treated in the next section.

3. Corpus Investigation

3.1 Methodology

This section investigates the *NP as we know it* expression (e.g., *the world as we know it*) and the same type of other expressions (e.g., *records as we knew them* and *his life as he knows it*) by employing the Corpus of Contemporary American English (COCA).[2] That is, this investigation focuses on a nominal modifier (i.e., *as*) with *know* specified as its clause's verb.[3] The reason for this is that the author's preliminary corpus research for the distribution of verbs that occur within the *as*-clause in question (e.g., *know*, *understand*, and *have*) shows that *know* is the most typical and frequent verb.

To retrieve data, the following search string is specified as the input: n* as * KNOW it/them/him/her/you/me/us. The signs *n** and * are used to denote any type of noun and a single arbitrary word, respectively. The verb *know* in capital letters includes all its variants (i.e., *know, knows*, and *knew*), and each oblique slanting line inserted between words means *or*.

3.2 Results and Observations

A total of 1580 examples with the nominal modifier *as* were collected from COCA. The first point to mention is that there is a great difference in frequency between two groups: group α, which takes *we* as the subject of the *as*-clause (1350 tokens), and group β, which takes other nominal phrases (e.g., *I, he, Adams,* and *people)* as the subject of the *as*-clause (230 tokens) (see Table 1). As for group α, the verb *know* is predominantly used as a present-tense form (1295/1350 tokens = over 95%). In fact, the string *NP as we know it* accounts for over 75% of all the examples (1197/1580 tokens). This bears out the prototypicality of this phrase vis-à-vis all the other variants (e.g., *NP as we knew them* and *NP as he knew it*).

Table 1 also shows information for the head nouns modified by the *as*-clause.

Table 1. Frequency and Varieties of Head Nouns Modified by the *As*-clause

group α (e.g., *NP as we know it*)		group β (e.g., *NP as he knows it*)
token frequency	1350 tokens	230 tokens
type frequency	435 types	86 types
varieties of modified nouns		
life (203), welfare (126), world (125), civilization (65), medicare (29), system (22), fact (18), industry (14), television (14), democracy (13), universe (13), internet (11), society (11), America (10), culture (9), stress (9), war (9), baseball (8), history (8), reality (8), security (8), education (7), science (7), capitalism (6), earth (6), government (6), IRS (6), party (6), process (6), story (6), time (6), church (5), country (5), landscape (5), law (5), music (5), politics (5), privacy (5), skiing (5)		life (90), world (28), truth (6), civilization (5), fact (5), television (5), family (4), universe (3), welfare (3), country (2), detail (2), discipline (2), marriage (2), medicare (2) (adolescence, America, bar, baseball, bird, Bulls, business, character, child, Christianity, circumstance, classroom, college, competence, culture, darkness, discourse, Espinoza, event, evidence, exhibition, face, field, future, history, hunting, internet, James, journalism, lifestyle, Macready, market, moderation, Momo, mountain)

As for nominal varieties, the number of occurrences is represented within parentheses. Because of space limitations, those that occur more than four times are listed in group α while those that occur more than once are listed in group β. Hapax legomena are partially indicated in group β. It can be seen in both groups that the same particular nouns (underlined in Table 1) frequently occur with the *as*-clause. Examples with these four nouns (i.e., *life*, *welfare*, *world*, and *civilization*) account for nearly 40% in group α and 55 % in group β. Their referents are considered closely related to living life, and they influence people all over the world. For example, the world is a place where we live, and welfare and civilization are necessary to enrich our lives. Without them, we would not be able to live a (normal) life, or at least our way of life would deteriorate. Moreover, most of the referents of the nouns listed in Table 1 do not assume a definite, concrete form, and they are difficult to imagine visually (i.e., abstract or conceptual). Another important point is that different types of nouns account for one-third of all the examples in both groups. Since type frequency generally correlates with productivity, this suggests that a variety of nouns are possible, although certain nouns, as mentioned above, are more likely to occur than others.

Furthermore, it was revealed that the referent of a nominal modified by the *as*-clause is very likely to take part in particular situations described by the matrix clauses (e.g., (1)) or nominal phrases (e.g., *the end of the world as we know it*). Through observation and data analysis, I found that the data can be classified into eight different types: "non-existence," "origination," "transformation," "correspondence," "non-concurrence," "existence," "other situations," and "non-situational." Examples of each category are offered below with its definition.

In the first place, examples belonging to "non-existence" are provided in (7).

(7) a. For if we fail, civilization as we know it will disappear.
(COCA, Magazine, 2014)
b. He has pledged to end welfare as we know it. (COCA, Spoken, 1996)
c. In Southeast Asian cultures, adolescence as Americans know it does not exist.
(COCA, Newspaper, 2000)

The situation "non-existence" is used to refer to a case where an entity (i.e., the referent of a nominal modified by the *as*-clause) is (going to be) absent or in crisis. A prototypical case found in the examples is one where an entity will cease to exist, be over, or be at stake, as in (7a-b). Its less-prototypical case is that an entity that exists in a certain area does not exist in another area, as in (7c). The difference between the two cases just lies in where an entity is absent: the former case indicates that an entity is (going to be) absent at some future point while the latter indicates that it is already absent in some area at the present time. Nevertheless, both cases share the fact that any entity corresponding to the one established and stocked in our mind cannot be found, whether in some area or at some future point.

Surprisingly, this type of situation accounts for nearly 50% (i.e., 793 tokens) of the entire examples. In other words, the referent of a nominal modified by the *as*-clause has a remarkable tendency to participate in this kind of situation. More precisely, this nominal has a strong preference for a hypothetical or future/past-oriented situation of "non-existence," as in (8a-b), over a present-oriented situation focusing on the lack of an entity at the time of the speech, as in (8c) (i.e., in the ratio of nine to one).

(8) a. If it [an asteroid] hits us, Earth as we know it will be over.
(COCA, Fiction, 1998)
b. Ten years ago, my world as I knew it ended. (COCA, Spoken, 2016)
c. "The software industry as we know it is dead," says Srivats Sampath, CEO of antivirus software maker [...]. (COCA, Magazine, 2002)

From these facts, it is not sufficient to say that the referent of the nominal in question tends to change, let alone to say that it is what is well-known now, as seen in (6a-b) in Section 2 (see also the ratio of "transformation" in Table 2 below). Instead, it is important to specify that its non-existence at a time other than the present is very likely focused on.

The following examples belong to the second type of situation "origination":

(9) a. Henry Steinway invented the piano as we know it. (COCA, Magazine, 2003)
b. The modern computer as we know it emerged in 1945, [...].
(COCA, Magazine, 2013)
c. "The Midwest as we know it began here." (COCA, Academic, 2006)

I use the term "origination" to represent a situation concerning the origin/birth of an entity (i.e., the referent of a nominal modified by the *as*-clause). Information about when/where it started to exist or who made it is typically involved in this case. For example, (9a) describes the invention of the piano and its inventor (i.e., Henry Steinway), (9b) the first appearance of the modern computer and its date, and (9c) the beginning of the Midwest and its place of origin.

What should be brought to attention is that these examples account for over 10% (i.e., 158 tokens) of the examples in group α while they account for only 1% (i.e., 3 tokens) in group β. The reason for this can be considered in the following way: information providing the origin of an entity becomes the most informative and beneficial when the entity is known by many people. If it were known to only one person, it would be futile to state its origin to others. In fact, the subject of the *as*-clause (i.e., the number of persons who recognize an entity) is plural in all the examples belonging to "origination" in group β, as exemplified in (10).

(10) […] 12.9 billion years ago, when the universe as humans know it was just beginning to emerge from the Big Bang. (COCA, Magazine, 2011)

That people all over the world know the presence of the universe is implied by the word *humans* in this case. Since the universe is known to everyone, information on its origin is regarded as informative and beneficial for hearers or readers, and thus worth noting.

The third type of situation is "transformation," whose examples are given in (11).

(11) a. Television as you know it is about to change. (COCA, Spoken, 2008)

b. Since liquid water is crucial to the evolution of life as we know it, the possibility of life on Mars does not stretch scientific credulity. (COCA, Magazine, 1997)

c. "This FreeMarkets auction idea," says Brittan, "is revolutionizing procurement as we know it." (COCA, Magazine, 2000)

The situation "transformation" is intended to denote a case where the form of an entity (i.e., the referent of a nominal modified by the *as*-clause) changes, or some aspect of it is altered for some reason, as in (11a). Cases such as (11b-c) are considered more specific instances of (11a), because *evolution* and *revolutionize* are words that each express some development (i.e., a certain type of change). This kind of situation accounts for about 10% (i.e., 160 tokens) of all the examples. We can see from this rate that "transformation" is not the first and foremost situational type when the *NP as we know it* expression or one of its variants is used.

The next type of situation "correspondence" needs some explanation. The term "correspondence" is used to represent a case where a person interprets two entities as being identical or compatible with each other. Strictly speaking, this type, unlike the three types seen above (i.e., "non-existence," "origination," and "transformation"), does not reflect a normal, objective event that seemingly occurs outside of a conceptualizer's (i.e., recognizer's) mind, but represents his or her inner activity, or cognitive process (i.e., how he or she construes the relationship between two relevant entities). For this point, I adopt the perspective of Cognitive Grammar (cf. Langacker (1987; 2008)), where any "objective" event cannot be recognized as such, but it is perceived via a conceptualizer's construal. What a person recognizes and describes necessarily reflects his or her construal. Therefore, I use the term "situation" in a broader way to cover any mental representation encoded in language.

Examples such as (12a-e) belong to "correspondence" and account for 4% (i.e., 64 tokens) of all the examples.

(12) a. […] the Middle Realm is the world as we know it; […].
(COCA, Academic, 1990)
b. Health insurance as we know it is illness and accident insurance, meaning there must be a diagnosis. (COCA, Magazine, 2004)
c. But the first guess is that it might be the cosmological constant, and that fits with the facts as we know them today. (COCA, Spoken, 2003)
d. […] trigonometry as we know it today is probably the result of Islamic religious rituals. (COCA, Spoken, 2002)
e. Santa Claus as we know him is a combination of stories from many different countries. (COCA, Spoken, 2005)

Examples (12a-b) intend to convey the referential identity between the first and second nominals. In (12c), each referent of the two nominals is construed as being compatible with one another. Examples (12d-e) directly convey the referential identity between the first and second nominals, but they also seem to imply how the referent of the nominal modified by the *as*-clause took shape. (12d) can be interpreted as "Islamic religious rituals contributed to the emergence of trigonometry" and (12e) as "combining stories from many different countries created a fictional character, Santa Claus"; nevertheless, both examples may not be purely considered to belong to "origination" due to the function of referential identity, but it would be safe to say that they are at least relevant to it. There are 15 examples of such a case. By classifying this situational type into two classes (i.e., (12a-c) and (12d-e)), we can find interesting syntactic behavior in each class. For the former class, 70 % of the nominals modified by the *as*-clause serve as the complement of the matrix verb while for the latter 73% of them serve as the subject of the sentence. This means that the nominals in question frequently function as ones specifying or identifying the (comparatively unfamiliar) subject referents in the former class whereas they are very likely to be used as topics in the latter. Their referents are already familiar to many people, as (6a) indicates, and therefore it is reasonable for these nominals to be used under the above conditions.

The fifth type of situation is categorized as “non-concurrence” and some of its examples are:

(13) a. Chanel Fluid Iridescent Eyeshadow ($30). This is not eye shadow as you know it. (COCA, Magazine, 2005)

b. Western democracy as we know it is incompatible with Zionism. (COCA, Newspaper, 2005)

c. On the positive side, the work would be different from journalism as she knew it […]. (COCA, Fiction, 1991)

This type of situation accounts for the same percentage as that of the previous one (i.e., 4%). Typically, it corresponds to a negated version of “correspondence,” as in (13a). I use the term “non-concurrence” to indicate a situation where by comparing an entity with another similar or relevant one, a person (i.e., conceptualizer) finds that they are incompatible with each other. Examples (13a-c) all convey the perceived difference between two entities, whether via the use of *not*, *incompatible*, or *different*. Interestingly, most of the examples belonging to this type (i.e., about 90%) take a nominal modified by the *as*-clause as the complement of the verb or preposition in the matrix clause, as in (13a) and (13c). This is similar to the former class in “correspondence” and further confirms that the nominals in question tend to provide more specific information for the subject referents in these two types (except for the latter class in “correspondence”).

Moreover, this type can be related to the situation “transformation,” where an entity changes and as a result it would not be (i.e., be incompatible with) the entity that it used to be. One difference between the two situations lies in the two referents considered incompatible with each other. They are the same entity in the case of “transformation” while they are typically not in “non-concurrence.” Another, important difference is that the latter situation does not depict any influence on the existence of the referent of the nominal modified by the *as*-clause. For this reason, the number of these examples is low compared to that of such situations as “non-existence,” "origination," and "transformation." The same thing can be said of "correspondence" as well.

Examples belonging to the situational type "existence" are exemplified in (14).

(14) a. Life as we know it exists in what I'll call the real world.
(COCA, Fiction, 2002)
b. You[']re going over there to save the world as we know it, son.
(COCA, Fiction, 2001)
c. […] liquid water on Earth's surface, which is generally agreed to be a prerequisite to sustaining life as we know it. (COCA, Magazine, 1996)

This type accounts for only 3% (i.e., 53 tokens) of all the examples. The term "existence" is used to indicate a situation where an entity (i.e., the referent of a nominal modified by the *as*-clause) is habitually present in some area or another entity keeps it existent. For example, the habitat for our familiar life is described in (14a), but many examples in this type do not just describe an entity's existence. Nearly 70 % of this type is used in situations that imply that an entity will be lost without some means or materials to keep it as it is. We can read from (14b) that the existence of the world is being in danger, which is implied by the meaning of the verb *save,* and the hearer enables it to exist, but it may cease to exist if he does nothing for it. Similarly, (14c) implies that our familiar life would not be able to exist without liquid water. In conformity with these implications, verbs such as *save*, *sustain*, *preserve*, and *survive* are used much more frequently than *exist*. Examples (14b-c), therefore, clearly have a connection with the situation "non-existence," and those simply representing an entity's existence (e.g., (14a)) are considered as extensions from the former and thus very low in number.

The remaining two categories are "other situations" and "non-situational," which account for 9% (i.e., 144 tokens) and 11% (i.e., 170 tokens) of all the examples, respectively. Some of these examples are shown in (15).

(15) a. And then Michael Deaver, one of President Reagan's closest aides discusses Nancy Reagan as he knows her. (COCA, Spoken, 2004)
b. I'd be happy to talk about the book, the writing process, life as we know it ...

(COCA, Magazine, 1996)

c. […]: the computer as we know it represents the world in digital bits—ones and zeros. (COCA, Magazine, 2001)

d. Comet C/2019 Q4 could not have formed in our solar system as we know it. (COCA, Magazine, 2019)

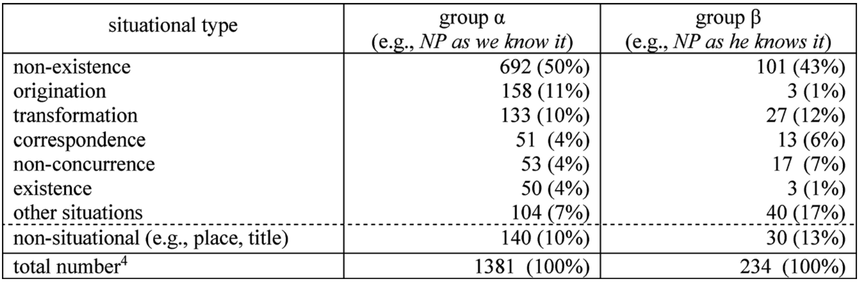
e. Bret’s most recent show, “Life As I Know It” ended with a proposal. (COCA, Spoken, 2011)

Examples (15a-c) are categorized into “other situations.” As the name suggests, this type consists of the set of examples that do not belong to any of the six types mentioned so far. For example, the situation of discussing or talking (e.g., (15a-b)) does not belong to any of them. Examples like (15c) are especially rare in that the referent of the nominal in question is considered to carry out something. Only five examples (i.e., 0.3%) are considered as members of this case. Importantly, it was also revealed that none of them depicts a situation where the referent in question causes another referent to change its state, which will be dealt with in more detail in Section 4.3.

Examples (15d-e) belong to “non-situational.” With the term “non-situational,” I intend to indicate that the referent of a nominal modified by the *as*-clause does not directly participate in the situation described by the matrix clause. For example, *Solar system* in (15d) serves as a place where the situation described by the matrix clause occurs and the underlined nominal in (15e) is the title of a show, which indicates that life itself does not take part in the situation described by the verb *end*.

Table 2 summarizes all the results of the distribution of the eight categories we have discussed so far.

Table 2. Distribution of Situational Types per Group



situational type	group α (e.g., *NP as we know it*)	group β (e.g., *NP as he knows it*)
non-existence	692 (50%)	101 (43%)
origination	158 (11%)	3 (1%)
transformation	133 (10%)	27 (12%)
correspondence	51 (4%)	13 (6%)
non-concurrence	53 (4%)	17 (7%)
existence	50 (4%)	3 (1%)
other situations	104 (7%)	40 (17%)
non-situational (e.g., place, title)	140 (10%)	30 (13%)
total number[4]	1381 (100%)	234 (100%)

Notably, the distribution of these situational types is highly skewed. The top three situations (i.e., "non-existence," "origination," and "transformation") account for about 70% of all examples in groups α and β. In particular, "non-existence" accounts for about 50%. As will be demonstrated in the next section, this strikingly skewed distribution results from the use of a nominal modified by the *as*-clause.

3.3 Comparison between Two Similar Types of Expressions

One may suspect from the results in Section 3.2 that expressions similar to *NP as we know it* also have almost the same results. However, this is not the case. To reveal this, this section compares the following similar, but different types of expressions: nominals modified by the *as*-clause whose subject and verb are *we* and *know* and ones by the *that/which* relative clause whose subject and verb are *we* and *know*. The search string of the latter is: *n that/which we KNOW. The total number of examples after removing noises and ambiguous cases[5] is 276 (tokens).

Among them, only 8 examples (i.e., 3%), including (16a), belong to "non-existence"; 10 examples (i.e., 4%), including (16b), belong to "origination"; and 3 examples (i.e., 1%), including (16c), belong to "transformation", all of which only account for 8 % of the total.

(16) a. A young lady that we know <u>died</u> last year from an overdose of pills.

(COCA, Newspaper, 2019)

b. The modern American society that we know today was just <u>beginning</u>.

(COCA, Newspaper, 2009)

c. […] and the standard model that we know now <u>will be</u> part of this larger model. (COCA, Spoken, 2001)

Table 3 shows a frequency comparison between *NP as we know it* expressions and *NP that/which we know* expressions, depending on four situational types (i.e., "non-existence," "origination," "transformation," and "others").

Table 3. The Cross Table of Two Variables "Expression" and "Situation"

	non-existence	origination	transformation	others	row total
NP as we know it type	692	158	133	398	1381
NP that/which we know type	8	10	3	255	276
column total	700	168	136	653	1657

I conducted a Chi-square test based on the data in Table 3, and the results revealed significant differences among conditions ($X^2(3)$ = 391.577, p < .001, V = 0.486).[6] Residual analyses also revealed that the *NP as we know it* expressions significantly prefer the three situational contexts "non-existence," "origination," and "transformation" and disprefer the context "others" while the *NP that/which we know* expressions significantly disprefer the former and prefer the latter.

Another important difference between the two kinds of expressions is that 7 examples (i.e., 2.5%) were attested where the referent of a nominal modified by the *that*/*which* clause plays a role of Agent or Causer in the matrix clause, as in (17), whereas their corresponding examples in the *NP as we know it* expressions only account for 0.3% of the entire examples.

(17) a. Even soloists that we know have <u>made such a claim</u>. (COCA, Magazine, 2000)

b. This gentleman that we know <u>flies Black Hawk helicopters</u>, so […]. (COCA, Spoken, 2003)

The agentive subject referent in (17a) does not affect the claim while that in (17b) changes the state of the helicopters. Remember the latter case was not attested in all the examples with the nominal modifier *as*. Consequently, these comparative data clearly show that a nominal modified by the *as*-clause has an idiosyncratic behavior of its own with respect to situational types its referent participates in.

4. Considerations on a Nominal Modified by the *As*-clause and Situational Types

4.1 The Basic Meaning of a Nominal Modified by the *As*-clause

The meaning of the phrase *NP as we know it* is not strictly predictable from its parts. Based on its descriptions in two dictionaries in Section 2 and observations on the corpus data in Section 3, it can be concluded that the pronoun *we* typically refers to not just interlocutors, but also many other people, which attracts words whose referents are well-known and felt close to them into the nominal slot. Moreover, the verb *know* specifically means familiarity with the present state of the referent of a nominal modified by the *as*-clause, rather than just recognizing it.

In addition to these aspects, especially important is the role of *as*. At the end of the last section, the findings revealed that this kind of phrase markedly tends to be more connected with "non-existence," "origination," and "transformation" than a similar phrase like *NP that/which we know* is. This skewed distribution can be accounted for as follows. Both connectors (i.e., *as* and *that/which*), whether directly or indirectly, function as restrictors on the range of a modified nominal, but it seems that the use of *as* relatively highlights knowledge of its referent at a certain point of time (reflected by the tense of the *as*-clause), as implied by the use of *now* in (6a) to specify the meaning of the *NP as we know it* expression, where the verb *know* is in the present tense. This can lead to the following typical situation: we have enough knowledge of a certain entity at a given point of time (typically, at the present), but we do not know how it will be or how it was. If we focus on the former, situations concerning "transformation" or "non-existence" will be chosen. If we focus on the latter, that of "origination" will be selected. Therefore, this kind of implication motivates the strong connection between the referent of the nominal in question and the three types of situations.[7] However, the reason why the case of an entity's disappearing is much more preferred to any other situation is not entirely predictable from any function of the nominal modifier *as* (and any meaning of the other elements used in this nominal). Therefore, these characteristics must be identified as idiosyncratic aspects of the nominal as a whole (see also the next section for a striking preference for "non-existence").

Given this implication, the same can be said for the other cases where subjects other than *we* are used in the *as*-clause (i.e., group β). The main difference between them is that the referent of the nominal in question is not necessarily known by many people in group β, as suggested in (18).

(18) The end of the world, at least the world as he knew it, had come.
(COCA, Fiction, 1998)

In (18), the nominal *the world* probably refers to the whole one known by everyone, but the nominal *the world as he knew it* does not correspond to the entire world per se, but to his personally experienced world; accordingly, other people do not necessarily know it. This difference leads to a further difference between the two groups with respect to "origination," as already discussed in Section 3.2.

4.2 The Relationship between "Non-existence," "Origination," and "Transformation"

The previous section has shown how the nominal modifier *as* plays a key role in the significant connection between the referent of the nominal in question and the three types of situations (i.e., "non-existence," "origination," and "transformation"). It is time to consider the relationship between these situations. They are argued to be typically subsumed under a common structured notion: a life cycle. That is, an entity, concrete or abstract, undergoes the following course of life: it is born, some of its aspects change as time goes on, and it finally becomes obsolete or extinct. If we consider a nominal and its modifying *as*-clause whose verb is *know* as a construction in the sense of Goldberg (2006), it follows that it evokes such a life cycle as the background knowledge needed to understand the phrase, or a frame in Frame Semantics (e.g., Fillmore (1982)). If this life cycle frame is so salient that it is entrenched and easily evoked as part of one's knowledge about the nominal construction in question, it would be reasonable to suspect that it induces the nominal to co-occur with the expressions depicting one of the three types of situations. Besides, since the notion of a life cycle involves existence of an entity, the nominal in question probably includes this more abstract notion as a secondary frame that thus motivates the occurrence of examples concerning "existence" as in (14). It is, however, the life cycle frame that is considered to be salient and easy to evoke. The existence frame's low saliency hampers its easy evocation, and hence the low number of such examples.

Among the three types of situations above, "non-existence" is outstanding in that a dominant number of examples belong to this type, as I have repeatedly mentioned. This fact may be accounted for in the following way. Constructs such as *NP as we know it* and *NP as he knew it* presuppose the existence of their referents, and therefore, their focus is more likely to be on how the referents will be than on how their existence was formed. Be that as it may, it remains an open question at this stage why the situation "transformation" is not as preferred as is the situation "non-existence." We need to further assume here that these constructs especially prefer to be used in a situation where their referents are going to undergo a radical, unfavorable change. The situation "non-existence" meshes well with such a change, because it typically describes cessation of existence (i.e., the transition from existence to non-existence) rather than an entity's minor changes. The situation "transformation," on the other hand, does not necessarily represent the same degree of change, as we can see in (11). From this line of reasoning, it follows that "non-existence" is particularly favored.

4.3 Functional Constraints on the Use of a Nominal Modified by the *As*-clause

The corpus investigation in Section 3 revealed that a nominal modified by the *as*-clause can be used with various situations, despite a strong preference for the three types of situations—"origination," "transformation," and "non-existence"—. Indeed, it seems that the use of this nominal is possible whenever a situation is compatible with its basic meaning discussed in Section 4.1.

However, the investigation also revealed that the referent of the nominal is very unlikely to participate in a situation where it carries out something, as in (15c). Consider the following relevant sentences:

(19) a. ??Albert as we know him <u>killed</u> three persons two years ago.
b. ??This gentleman as we know him <u>flies</u> Black Hawk helicopters.

Example (19a) depicts a situation relevant to "non-existence" and (19b) corresponds to example (17b), but neither of them is accepted as natural. Here each subject referent causes each object referent to change its state. Remember no such cases were attested in the corpus investigation. Then, the following constraint on the use of the nominal in question is in effect: the referent of a nominal modified by the *as*-clause must not cause another entity to change its state. Why does such a constraint exist?

Huddleston and Pullum (2002: 1150) provide an insightful statement about the nominal modifier *as*. They point out that the *as*-clause in question specifies some property or aspect of the referent of the modified head noun.[8] If each of the subject referents in (19) denotes some or a set of properties that constitute the individual, these marginal sentences are naturally accounted for. Specifically, the properties of an entity are unlikely to be interpreted to influence another entity in these cases. What causes the object referent to be killed or flown in (19a-b) is the person, not his properties themselves. Because of the presence of the *as*-clause, these sentences sound unnatural.

Having set the stage, let us now go back to (15c), repeated as (20).

(20) […]: the computer as we know it represents the world in digital bits—ones and zeros. (COCA, Magazine, 2001) (= (15c))

Example (20) is similar to (19) in that the subject referent carries out some action, but there are two points to note. The first point is that the subject referent in (20) does not affect the object referent's state. It just computationally describes the world. The second one is that this sentence does not indicate any individual situation that happens at a particular time. Rather, it presents a general fact. Since it is predicated of one of the intrinsic properties of the computer, it is regarded as acceptable and these examples are attested.

With this in mind, consider the following example[9]:

(21) Gender diversity as we know it today will destroy women's sports in the future.

This example will be judged acceptable even though the nominal modified by the *as*-clause serves as the subject whose referent causes other entities to be destroyed (i.e., to change their state). Two aspects make this sentence acceptable. The first point is that this sentence probably depicts a situation expected to happen at any point of time in the future, not one that actually happened as in (19a). This means that destroying women's sports can be considered as one of the stable properties of gender diversity, which is compatible with the explanation given for (20). Another point is that the nominal phrase *gender diversity* encodes an abstract notion (not a concrete or animate entity) and represents some property in itself. Hence it has compatibility with the above-mentioned function of the nominal modifier *as*. Nevertheless, it is crucial to say that most of the attested examples are concerned with what happens to the referent of a nominal modified by the *as*-clause, not what it does. This overall tendency, therefore, disprefers cases such as (19)–(21) to occur in the actual examples.

There is another important condition imposed on the nominal in question. We already saw in (12a) that it can function as the complement of the predicate in the matrix clause and provide more specific information for the subject referent. This conforms to its meaning, since many people are likely to be more familiar with the former than the latter. This nominal, however, requires careful attention when it functions as the subject in a sentence. Specifically, this nominal cannot be used in a case where a well-known fact is predicated of the referent of the nominal to specify the kind of entity or property it belongs to:

(22) ??Paris as we know it is a beautiful city.

This can be improved if the predicate is adapted with the phrase *is not just a beautiful city, but has a dark side* or if it is simply replaced with the phrase *has a dark side*. Either phrase implies that the attributive or identificational information introduced for a nominal modified by the *as*-clause needs to be unexpected or unfamiliar to hears or readers (at least at the speech time).[10]

To sum up this section, at least two conditions are needed to reflect characteristics of the attested examples: (a) the referent of the nominal in question must not cause another entity to change its state unless its act is regarded as some property it has and (b) the attributive or identificational information predicated of this referent needs to be unexpected or unfamiliar to hearers or readers.

5. Conclusion and Outlook

This study confirms the collocational preference of a nominal modified by an *as*-clause whose verb is specified as *know*, and its basic semantic and functional properties. As for the former, I revealed that the referent of the nominal in question significantly tends to participate in one of the situations describing its "origination," "transformation," and "non-existence." Among them, the situation "non-existence" is the most typical, accounting for half of all examples with this kind of nominal in COCA. Classification of this nominal into two groups also uncovered the fact that group α is an unmarked, typical variant (e.g., the string *NP as we know it* accounts for 75% of all the examples). Group β greatly differs from group α in that the former includes very few examples expressing the situation "origination" (i.e., only 1 %). I claimed that information about an entity's origin is beneficial when the entity is known by many people, as in group α, and is futile when only a single person knows it. Group β includes many examples where it is only known by a single person or a small group and hence disprefers this type of situation.

With respect to the basic meaning of the nominal in question, one must be aware of the fact that the subject referent of the *as*-clause feels familiar with the referent of this nominal. Importantly, the use of this nominal is claimed to convey that its referent is well-known at a certain point of time (typically, at the present), but it is unclear how it will be or how it was. This implication is considered to motivate the referent's significant tendency toward participating in the three types of situations above. The assumption that a situation where an entity is going to undergo a radical, unfavorable change is particularly preferred by this nominal further accounts for a predominant number of examples concerning "non-existence."

In regard to restrictions imposed on this nominal, I revealed two functional conditions. The first one is that the referent of the nominal must avoid causing another entity to change its state unless its act is regarded as some property it has. The second one is that the attributive or identificational information predicated of the subject referent of the nominal needs to be unexpected or unfamiliar to hearers or readers. Thus, cases are ruled out in which something self-evident is predicated of this referent.

This study examined and revealed the nature of a specific type of construction: an *as*-clause whose verb is specified as *know* modifies a preceding nominal expression. To clarify and cover the full scope of the nominal modifier *as*, however, there is much work to be done from both synchronic and diachronic perspectives. As the next step, we need to conduct a synchronic investigation into what types of verbs are allowed to occur in the *as*-clause (e.g., *life as we understand it* and *reality as we have it*) and what types of situations tend to be described. After uncovering these points, we can postulate general functions of the nominal modifier *as*. From a diachronic perspective, unveiling the historical development of this kind of nominal expression is crucial—specifically, how the *as*-clause came to modify the preceding nominal expression. It is also important to ascertain what led to the present skewed distribution of situational types revealed by this study. In this sense, this paper has significant importance as a foundational study toward clarifying the full scope of the nominal modifier *as*.

Acknowledgements

An earlier version of this paper was presented at the 28th annual meeting of the Society of English Grammar and Usage held online, October 17–23, 2020 and I appreciate the comments offered by participants in this meeting. I am very grateful to Kazumi Taniguchi, Takeshi Omuro, Seizi Iwata, Takashi Mino, Takuya Hirata, Shimon Nakanishi, Tetsuya Nakamura, and all the members of the cognitive linguistics laboratory at Kyoto University. I also thank four informants for continuously and patiently cooperating with me and three anonymous reviewers for their helpful comments and suggestions. This study is supported by JSPS KAKENHI GRANT NUMBER 21J12310. All remaining errors are, of course, my own.

Notes

1. None of these dictionaries does no more than cite an example of situations describing some change:
   (i) This could mean the end of life as we know it. (MEDAL2)
2. The first access to the data was dated January 23rd, 2020, and COCA was updated in March 2020 to expand its scale, so this paper does not include those data that belong to "TV/MOVIES" and "BLOG."
3. There also exists the following type of sentence:
   (i) "Live TV, as we know it, is over," she says. (COCA, Magazine, 2004)
   However, this (marked) variant with commas occurs much less frequently (i.e., 1/15 times) than the other (unmarked) variant without commas. As this paper aims to reveal the basic properties of the *NP as we know it* expression and the same type of other comma-less expressions, these unmarked cases are dealt with in the following sections. I leave their differences to my future research.
4. The total number in this table is different from that in Table 1, because the referent in

question can participate in more than one situation when the two (or more) predicates are conjoined within the matrix clause in a sentence:

(i) Does all this mean that heavy metal music as we know it is dead, buried and beyond resuscitation? (COCA, Newspaper, 1997)

The number of situations is counted within the range of a matrix clause in cases like (i). If there is any noun depicting some situation as in (ii), the range is restricted to such a nominal phrase, which includes its head, an *as*-clause, and another nominal modified by this *as*-clause.

(ii) […] we're finally going to see the end of welfare as we know it. (COCA, Spoken, 1996)

In (ii), the noun *end* is considered to denote a situation and thus the situation depicted by the verb *see* is not counted.

5. The following case is ambiguous whether the relative clause serves to modify the preceding nominal or restrict a range of the speaker's knowledge (as when using the phrase *as far as we know*), according to Kono (2012):

(i) This is not explicable by any means that we know. (COCA, Spoken, 2007)

He points out that in the latter case, certain elements such as negatives (e.g., *no*), numerals, or quantifiers (e.g., *all*, *every*, *any*) generally accompany the modified nominals. Therefore, this paper excludes examples with these elements. However, even if they are included in the total number, the results are much the same as those without these examples.

6. To analyze the data, I used js-STAR, a data analysis software, which is available online at https://www.kisnet.or.jp/nappa/software/star/.

7. It is, of course, possible that this implication is not overtly reflected by any linguistic material, especially when the referent of the nominal does not directly take part in a situation (i.e., in the case of "non-situational").

8. Although they do not discuss why the nominal modifier *as* has such a specification in detail, the following point is suggestive for it. Huddleston and Pullum (2002: 1150) (and also Yagi (1996: 219)) point out that the nominal in question (e.g., *the world as we know it*) can be paraphrased as a sentence with *as such and such* (e.g., *We know the world as such and such*). There are similar phrases like *regard the situation as serious*, where the adjective *serious* depicts one (potential) property of this situation. In the same vein, the part *such and such* in *We know the world as such*

and such is analyzed as depicting some property of the entity functioning as the complement of the matrix verb and this part corresponds to *the world* in *the world as we know it*, as a result of which this nominal can be considered to depict some property associated with it. The detailed analysis will be left to my future research.

9. I owe this example to an anonymous reviewer in the Society of English Grammar and Usage.

10. This condition does not seem idiosyncratic to the nominal of this type, because a similar sentence (e.g., ??*The Paris that we know is a beautiful city*) is also unnatural. I leave the scope of its application for future research.

References

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

Fillmore, C. J. (1982) “Frame Semantics,” In The Linguistics Society of Korea (eds.), *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Company, 111–138.

Fukumura, T. (1985a) “Daimeishi no Umu to As no Hinshi,” *Eigo Kyoiku* [The English Teacher’s Magazine] 33 (5), 58–60.

Fukumura, T. (1985b) “Daimeishi no Umu to As no Hinshi (Sono 2),” *Eigo Kyoiku* [The English Teacher’s Magazine] 33 (7), 64–65.

Goldberg, A. E. (2006) *Constructions at Work: The Nature of Generalizations in Language*. Oxford: Oxford University Press.

Hirota, N. (1988) “As-setsu to Kankeishi,” *Eigo Seinen* (The Rising Generation) 133, 9.

Huddleston, R., and G. K. Pullum (2002) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Kanaguchi, Y. (1978) *Gendai Eigo no Hyougen to Gokan*. Tokyo: Taishukan.

Kinugasa, T. (1979) “Toki wo Arawasu As,” *Gohou Kenkyu to Eigo Kyoiku* (Studies in Usage and English Teaching). 1, 17–25.

Kono, T. (2012) *Eigo no Kankeishi*. Tokyo: Kaitakusha.

Langcker, R. W. (1987) *Foundations of Cognitive Grammar, vol. 1: Theoretical Prerequisites*. Stanford: Stanford University Press.

Langcker, R. W. (2008) *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Ogawa, A. (1985) “Meishi-ku wo Gentei Suru As-setsu,” *Eigo Kyoiku* [The English Teacher’s Magazine] 131, 444–445.

Yagi, K. (1996) *Neitibu no Chokkan ni Semaru Gohou Kenkyu: Gendai Eigo heno Kijyutsuteki Apurouchi*. Tokyo: Kenkyusha.

Dictionaries

Collins COBUILD Advanced Learner's Dictionary, 9th ed. (2018) Glasgow: Harper Collins Publishers.

Macmillan English Dictionary for Advanced Learners, 2nd ed. (2007) Oxford: Macmillan Education.

（佐藤　嘉晃　京都大学大学院）

「論文」

日本国憲法をめぐる日本の帝国議会・国会会議録における「翻訳」という語の使用比較分析—翻訳調憲法の受容の変遷

島津美和子

Abstract

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

1. はじめに

日本国憲法施行から75年が経過した現在，日本政府が進める改憲の理由の一つに，現行憲法はGHQから提示された英語原案を日本語に翻訳したものであり，日本が策定したものではないという主張がある[1]。例えば，自民党が2015年に発表した政策パンフレット『ほのぼの一家の憲法改正ってなあに？』では,孫の「憲法って変な日本語が多くない？」という疑問に対し,祖父が「今の憲法の前文にはアメリカ合衆国憲法やリンカーンの演説なんかが翻訳口調で

使われているからなぁ」「日本国憲法の基（もと）[2] を作ったのがアメリカ人だからじゃよ」と答える場面がある。祖父の主張は，換言すると，現行憲法の原案はアメリカ人が作成したもので，それを翻訳した現行憲法は翻訳の口調が残り，質が悪いということだ。ここでは「翻訳」という語が否定的な意味で用いられている。憲法が「翻訳口調」であるゆえ，悪文だという祖父の見解は，一般的なものなのだろうか。そこで，本研究では，まずこの見解が 1946 年当時，憲法改正草案を審議していた当時の帝国議員らの発言や昨今の憲法改正論議の国会議員の発言にもみられるか否かを検証するため，帝国議会・国会会議録をコーパスとみなし，帝国議会・国会において議員が帝国憲法・日本国憲法を討議する際の翻訳[3] の用例を抽出し，形式的・意味的な特徴を分析する。参考事例として，GHQ の主体となった米国の議会の会議録内の憲法の文脈における translation の用例を抽出し，同様に分析する。これらの分析結果を比較し考察を行い，結論を提示する。

2. 先行研究

最初に辞書におけるこれらの翻訳関連の語の語義を確認しておく。文章を評価する際に用いる翻訳の複合語として，サ変接続名詞「翻訳」[4] と名詞的接尾辞「調」から成る「翻訳調」とサ変接続名詞「翻訳」と普通名詞「口調」から成る「翻訳口調」がある。「翻訳口調」はいずれの国語辞書でも見出し語になっていないが，「翻訳調」については一部の国語辞書に立項されている。日本最大規模の国語辞書とされる『日本国語大辞典』には立項されておらず，立項している辞書の語義は『広辞苑』の「外国語を日本語に訳してできたような，それまで日本語では使わなかった表現や文体」（p. 2733）と『大辞林』の「外国語を日本語に直訳したような独特の表現。また，そのような文体の作品」（下，p. 2407）である。このことから, 2 つの辞書の辞書編纂者は少なくとも「翻訳調」の語自体に否定的な意味を付与していないことが分かる。ただし，翻訳調の語釈は，具体的にどのような表現をもって外国語を日本語に翻訳した文章の特徴とみなすのかについては答えていない。辞書の利用者の解釈に委ねられている。一方で，「翻訳調」を構成する具体的な表現を抽出して「翻訳調」の概念を説明する言語学者もいる。例えば，石黒（2007: 46–64）は，「無機質な感じ」「連体修飾表現」「英語の翻訳を想起させるような表現選択」の 3 つを，また，大岡（2017: 15）は，人称代名詞・無生物主語，関係代名詞，比較級，動詞進行形，

受動態・使役の5点をその特徴として挙げる。石黒は「連体修飾表現」を除いて読み手が受ける主観的印象をもとにしているが，大岡の挙げる特徴は客観的に判別可能な文法的機能となっている。石黒の「無機質な感じ」は大岡の無生物主語，「連体修飾表現」は関係代名詞，「英語の翻訳を想起させるような表現選択」は比較級，動詞進行形，受動態・使役にそれぞれほぼ対応し，両者の間には共通項があると考えられるものの，翻訳調はさまざまな捉え方が可能で，厳密な定義は難しい概念といえる。

上記を念頭に先行研究を研究テーマと研究データの2つの観点から述べる。

第一に，「翻訳調」を主要テーマとする論考は，タイトルまたはキーワードに「翻訳調」を含むものに絞ると，日本語文献には数少なく，国文学と通訳翻訳研究で質的研究が各1件ある（伊原，2000; 梅林，2000）。しかし，これらは文学を扱っており，本研究のように法令文を対象としていない。

一方，英語文献では，「翻訳調」の英語相当表現として，名詞のtranslation, translatorおよび動詞translateに接尾辞-eseを付加した名詞のtranslationese, translatorese, translateseがある。「翻訳調」と同様にこれら3語の語義を辞書で確認する。『オックスフォード英語大辞典』（OED）はこれら3語に同一の語釈を当てており「（悪い）翻訳の性質と感じられる言語様式。翻訳文書の中のぎごちない，不自然，または，慣用的でないようにみえることば（language），特に翻訳者が原文に特有の特徴を訳文にできるだけ持ち込むように試みた結果，生じたことば」（以上，筆者訳。以下，原典の訳は筆者による）としている。否定的な意味が認められる。また，OEDは接尾辞-eseの現代的な用法として批判する対象の文章に付加することを説明し，例としてnewspaperese, novelese, officialese, journaleseを挙げる。つまり，「翻訳調」の「調」と違い，接尾辞自体に否定的な意味がある。一方，米語辞書の『メリアム・ウェブスター英語辞典』（MWD）には，translatoreseのみが立項され，その語義は「翻訳者に特有の用語（ジャーゴン），訳質の悪い文書」となっている。また接尾辞-eseについては「話し言葉，特定の場所，人，集団，学問，題材，活動に特有な話し言葉，書き言葉，または言い回し」であり，「通常，侮蔑に用いた語についていう」と説明している。OEDの語釈と同様，単語と接尾辞の両方に，否定的な用法を注記している。つまり，米語の場合も，translatoreseや接尾辞-eseが否定的な意味を持つことが分かる。また，見出しの選択についてはOEDとMWDで違いがある。実際の頻度を，書籍の頻度をもとにしているGoogle Books Ngram Viewer[5]でみると，translationeseが他の2つ（translatoreseとtranslatese）と比べ

圧倒的に多い。

以上のことを踏まえ，translationese をタイトルまたは要約に含む論考をみると, 1990 年から 2021 年の 30 年間に 1 件あるかないかの程度で推移してきたが, 2000 年頃から 2〜4 件となる年が増えてきている。translationese は元の言語で書かれた文書（非翻訳文書）と翻訳された文書（翻訳文書）とを区別する指標として用いられてきたが，近年は文書自動分類システムの開発を視野に入れた研究がみられるようになったことがこの増加と関係していると思われる。また，論文の被引用数（Google Scholar による）については 1 桁から 3 桁台まであり，ばらつきがある。被引用数が 200 を超える論考は 2 件あり，多い順に Baroni and Bernardini（2006），Tirkkonen-Condit（2002）となっている。いずれも量的研究である。簡単に内容を紹介すると，Baroni and Bernardini（2006）は機械学習を用いて translationese を同定するアプローチについて考察したものである[6]。一方，Tirkkonen-Condit（2002）は，「translationese は存在するのか」という問いを設定し，フィンランド語コーパスを使って人間の被験者に非翻訳記事と翻訳記事とを区別できるかを実験し，考察を加えた論考である。分析結果は，非翻訳記事と翻訳記事それぞれの言語的特徴は明示的には区別の手掛かりになっていなかったというものだった。これは，文書のジャンルや翻訳の質の方が非翻訳記事と翻訳記事の差異よりも特徴的にあらわれていたためであった。例えば，被験者は翻訳品質が高いと翻訳記事を非翻訳記事と誤って判別する傾向があった。

translationese の文献のうち，日本語について論じた論考には文学を対象とした 3 件（Meldrum, 2009a, 2009b, 2009c）と翻訳調の受容度に関する調査報告 1 件（Furuno, 2005）がある。同調査報告は翻訳者を対象に翻訳調の逐語的な日本語と自然な日本語のどちらを好むかを調査したもので，後者の率が高いことを報告している。

なお,「翻訳調」の英語相当表現を translationese としたが，これらは厳密には等価とはいえない。翻訳そのもののあるべき姿を論じた Newmark（1991）は, translationese とは翻訳者の母国語文書を外国語に訳した場合にみられる事象であり，translationese が悪い理由は，事実を言い表していない，文法的に不適切なためでは決してなく，原文のトーンやムードが訳文では正しく伝わらず，ぎこちないためだと述べている（pp. 21–22）。続いて「干渉の美徳と翻訳調の悪徳」（The Virtues of Interference and the Vices of Translationese）という章を設け, translationese を干渉（interference）と対比させながら説明する。両者とも起点

言語の影響により生じた現象であることは共通しているが，translationese が誤りであるのに対し，干渉は誤りでない点において両者は異なるとしている（p. 78）。そして translationese は無知または注意力の欠如による誤りであり，直訳では意味がずれたり，曖昧になるか，明確な理由もなく目標言語の用法に反する現象だと説明する（p. 78）。つまり，翻訳を行う側にとって translationese は避けるべきものということである。

第二に，国会会議録を「戦後 60 年近くにわたる…話し言葉での膨大な発話記録」（松田，2008: 1）あるいは大規模な日本語データ（松田，2012: 55）と捉え，言語研究に生かすことがなされてきた。松田（2008）では，同会議録を生かした語彙論，形態論，談話分析，社会言語学，方言学，自然言語処理といった言語学の各分野の研究が紹介されている。ここ 2～3 年の文献数は 2～3 件であるが，2015 年から 2019 年までは 4～7 件の範囲で推移していた。国会会議録検索システムが 2014 年から Web API として提供されるようになり，利便性が増したことも関係していよう。研究内容としては，国会会議録が第 1 回国会（昭和 22 年 5 月）から現在に至るまでの会議録であることを利用して，ある特定の表現の使用実態と変遷を追い，その要因を言語的に分析する通時的研究が多い傾向にある。例えば，浅川（2019）は，形態的に「ら」が過剰に使用されている用例の出現箇所を調査し，1950 年代の使用率が最も多いこと，また 1950 年代の傾向と 2000 年代の傾向には形態的に違いがあることを見出している。一方，帝国議会会議録を言語研究に用いた論考はこれまで 1 件にとどまっている。本論考のように帝国議会会議録と国会会議録の双方を言語研究に使用した例は管見の限りこれまでない。

3. リサーチデザイン

3.1 研究目的と研究設問

本研究は, 現行憲法は翻訳の口調が残り, 質が悪いという見解が 1946 年当時, 憲法改正草案を審議していた当時の帝国議員や昨今の憲法改正論議の国会議員の発言における表現上の特徴にもみられるか否かを検証することを目的とし，研究設問を「日本の憲法の文脈において『翻訳』という言葉について国会の場でどのように語られてきたか。また，日本国憲法成立後と成立前で語られ方に違いがあるか」と設定する。会議録は一つのジャンル，すなわち，「メンバーがコミュニケーション上の諸目的を共有するコミュニケーションの事象の一

種」（Swales, 1990: 58）を構成すると考えられる。Swales によれば，これらの諸目的はジャンルの存在理由をなし，その結果，ディスコース構造を規定し，内容とスタイルの選択に影響を及ぼすとし，さらには，それらの内容とスタイルには特定のパターンがあるとしている（1990: 58）。ジャンル分析と関連の深いものにムーブ分析がある。ムーブとは同じく Swales が考案した研究枠組みであり，端的には「ジャンルのコミュニケーション上の全体的目的を達成するテキスト中の機能的単位」（Kanoksilapatham, 2007: 24）であり，あるジャンルの典型的なテキストはこの一連のムーブから構成されている（Kanoksilapatham, 2007: 24）。ムーブ分析は，幅広い学問分野と専門分野のディスコースの研究に適用されているが，会議録に焦点を当てたムーブ分析の研究の論考は特定できていない。しかしながら，ムーブという用語を使わなくともそれに類する研究はあり，例えば，山口（2017）では，国会会議録のうち，衆参両議院の本会議・予算委員会の会議録に限定した場合, 会議は (1)「趣旨説明」, (2)「質疑」, (3)「討論」，(4)「採決」の 4 段階があるとしている。本研究が対象とする会議録もこれと同じ段階を踏むが,「翻訳」が語られる場面は質疑と討論が中心となる。いずれも，質問者，回答者，発言者は聞き手に自らの主張を理解してもらい，賛同を求めるため，いわば正当化するために質疑，回答，発言を行うと考えられる。そこで本研究では，そうした目的のために，話者がどのような特徴の表現を用いているかに着目する。

3.2 データ

3.2.1 日本語データ

帝国議会会議録および国会会議録から成る。前者については，帝国議会会議録検索システム（https://teikokugikai-i.ndl.go.jp）に対し，キーワードに「帝国憲法改正 AND 翻訳」[7] を設定し検索し得られた 36 会議の会議録のうち，憲法の文脈で「翻訳」を用いた発話を日本語データに使用する（付録表 1)。付録表 1 に示す会議録は帝国議会における帝国憲法改正案の審議の過程を記録したものであり, これらの会議録内の発話は帝国憲法改正案をめぐり議員が「翻訳」をどのように語っているかをみるために使用する。衆議院の帝国憲法改正案委員会（芦田均委員長）および帝国憲法改正案委員小委員会（芦田均委員長，通称芦田小委員会)，貴族院の帝国憲法改正案特別委員会（安倍能成委員長）の会議録が大半を占める。これらの会議では，各議員から出された意見をもとに枢密院にて可決された憲法改正草案が審議され，草案に各種の追加や修正がな

された。

一方，国会会議録については，国会会議録検索システム（https://kokkai.ndl.go.jp）に対し，キーワードに「（翻訳調 OR 翻訳臭）AND 憲法」[8] を設定し検索した結果，得られた会議のうち，憲法の文脈で「翻訳調」または「翻訳臭」を用いた発話を日本語データとして使用する（付録表 2）。文字列としては「翻訳調」と一致するが「翻訳調査」など「翻訳調」を意味しないものや刑事訴訟法など憲法以外の法令を指して「翻訳調」「翻訳臭」と述べているものは除外してある。付録表 2 内の発話は現行の日本国憲法と翻訳の関係について議員がどう語っているかをみるために使用する。検索対象を「翻訳」ではなく「翻訳調」「翻訳臭」に限定したのは，帝国議会会議録でまとまった数，生起していた「翻訳調」，「翻訳臭」に限定することによって，目視で分析可能な数に抑えるためである。

3.2.2 英語データ

アメリカ議会図書館の運営する Congress.gov で公開されている 56 議会（1899-1901）～103 議会（1993-94）および 104 議会（1995-）以降の *Congressional Record* に対し，“Constitution translation” および “translation of * Constitution”[9] を検索語とし，当該サイトの検索システムで検索し得られた結果を 20 会議分の英語データとして使用した。検索時，検索語はその派生語も含めて検索するよう条件設定した。なお，translationese，translatorese, translatese の用例は文脈に関わらず皆無であった。

3.3 手法

3.3.1 データの事前処理

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

3.3.2 分析の手順

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

4. 結果と考察

4.1 帝国議会会議録

帝国議会会議録中の「翻訳」を発言目的別に分類した結果を表 3 に示す。翻訳の行為ないし翻訳結果を指し示す場合が約 8 割を占め，この場合，発話者は翻訳に対して中立的であった。批判に用いる場合は，「翻訳調」（1 件）よりもむしろ「翻訳（の）口調」（11 件）と表現することが多かった。「翻訳（の）口調」の例を（1）に示す。

（1）「飜譯の口調[10] があつて，而も餘り巧みな飜譯であると云ふことは申されぬやうであります，細かい修正意見は委員會に於て申しまするが，私は洵に此の文字の用ひ方に於て遺憾を感じて居ります，條文は簡單でありまするから，さう飜譯口調が出て居ると云ふやうな所がありませぬが，此の前文には飜譯口調が大いに出て居ります,」（日本自由党　北れい吉　第 90 回帝国議会　衆議院　本会議　第 5 号　昭和 21 年 6 月 25 日）

表 3. 帝国議会会議録中の「翻訳」

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

注 1　会議録は，旧字体を用いているが上記表では適宜新字体にあらためた。

注 2　「実際の表現」列の括弧内の数字は件数を示す。

ただし，これには注意が必要である。表 3 の各表現の件数は，発話者の発話の回数に左右されるためである。実際，それぞれの発話者は批判のために使用する表現が異なっていた（表 4）。複数の発話者に使用された表現は「翻訳口調」と「翻訳的」にとどまっており，批判のための統一的な表現はなかったことが分かる。

表 4. 憲法を批判するために使用した表現（発話者別）

発話者	所属会派	批判に使用した表現
山本勇造	無所属倶楽部	翻訳臭
廿日出厖	日本自由党	翻訳的
森戸辰男	日本社会党	翻訳の印象
北れい吉	日本自由党	翻訳口調，翻訳の口調，翻訳的
山田悟六	日本進歩党	翻訳文の感
原健三郎	日本進歩党	翻訳口調
安部俊吾	日本自由党	翻訳の臭ひ

また，特に厳しく批判する場合は，「極めて」「甚だ」「全く」といった程度副詞と共起していた。表 4 の表現は，品詞的には名詞句および形容動詞があるが, 名詞句の場合は形容詞や形容動詞を付加した上で程度副詞を用いている（例(2)）。ここでは「甚だ ... 濃厚」と形容することで，翻訳臭さが存在するだけでなく，発話者が翻訳臭いと評価する基準点を憲法の翻訳臭さの程度が上回っていることを意味する。つまり，より強く批判していることになる。『使い方の分かる類語例解辞典』においても「甚だ」は「普通の程度を超えているさまを表わし，特に自分にとってマイナスの物事の場合に多く使われる」（p. 1005）と注記されている。

(2)「私は本草案が，既に北，鈴木兩同僚に依つて指摘された如く[11]，甚だ飜譯の臭ひが濃厚であると思ふのであります,」（安部俊吾　無所属倶楽部　第 90 回帝国議会　衆議院　本会議　第 8 号　昭和 21 年 6 月 28 日）

批判のための表現は，発話者本人は必ずしも批判の立場に立っておらず，中立を保ちつつ，他者の批判に言及する際にも件数は 2 件にとどまるが用いられていた。例（3）では，憲法の調子が翻訳的であると評することは非難に相当

すると明言している。

(3)「又新憲法の調子が甚だ飜譯的であると云ふ御非難もありましたが，匆々の際，或は用語等に於て御不滿な所もありませうが，是は十分に御審議を下すつて，御滿足の行くやうに御改めになられたいと思ひます[12]」（吉田茂　内閣総理大臣兼外務大臣兼厚生大臣臨時代理　第90回帝国議会　衆議院　本会議　第5号　昭和21年6月25日）

4.2 国会会議録

国会会議録中の「翻訳調」「翻訳臭さ」を発言目的別に分類した結果を表5に示す。

表5. 国会会議録中の「翻訳調」「翻訳臭さ」

発言目的	件数	程度副詞類が前接する件数	強意表現が前接する件数
発話者本人が憲法前文または憲法全体を批判するため	22（うち1件は「翻訳臭さ」）	5	0
他者の憲法前文または憲法全体に対する批判に言及するため	9	1	0
発話者本人が自党を代表して憲法前文または憲法全体を批判するため	6	1	1
延べ数	37	7	1

なお，名詞句「翻訳臭」は国会会議録にはなく，名詞「翻訳」に形容詞型の接尾語「臭い」を付加した「翻訳臭い」の連体形の名詞化「翻訳臭さ」が1件あるにとどまった。発言目的別にみると発話は発話者が自ら憲法を批判するケースが過半数を占め，発話者が自党を代表して批判するケースと合わせると7割強になる。帝国議会と異なり（表4の所属会派の列参照），発話者の属する政党によって，日本国憲法の文章に批判的かどうかが分かれていた。発話者自身が批判するケースと自党を代表して批判するケースの合計件数の2割強は，「極めて」，「非常に」，「余りに」，「余りにも」の程度副詞（例（4）），ある

いは，名詞「切り」に由来する成句「切りがない」に副助詞「くらい」を伴った「切りがないくらい」という程度を示す表現を用い，手厳しく批判していた(例 (5))。あるいは名詞句としての「翻訳調」を強意表現「全くの」で修飾することによって，批判の程度を強めていた。

(4)「現行憲法を見てみますと，どうでしょう。制定の経緯から，非常に翻訳調でわかりにくい表現が多くあります。」（近藤三津枝　自由民主党　第180回国会　衆議院　憲法審査会　第1号　平成24年2月23日）

(5)「憲法前文から日本語の表現としていかがなものかというようなことを取り上げれば切りがないくらい翻訳調で，国語としてなっていないというふうに思っております。」（山谷えり子　自由民主党　第179回国会　参議院　憲法審査会　第3号　平成23年12月7日）

また，批判する発話者は「翻訳調」を補完する目的で，しばしば言い換え表現を使っていた。「わかりにくい」が複数あり，このほか「日本語にややなじまないような語感を持った」，「悪文」，「美しくない」，「日本語として違和感がある」があった。これらの表現は，各発話者が具体的にどのような意味で「翻訳調」を用いているかを示す。発話者によっては，現状の憲法がどのような点で「翻訳調」であるかではなく，それをどのような表現に改めるべきかを説明している。この場合，そうした表現の否定がその発話者にとっての「翻訳調」を示すと考えられる。例えば，(6) の発話者は「翻訳調」とは難解であり，わかりにくく，模範的な日本語ではないと捉えていると推察できる。

(6)「前文の文章表現に関しましても，1. 翻訳調の現行の前文の表現を改め，前文の文章は，平易でわかりやすいものとし，模範的な日本語の表現を用いるべきである」（保岡興治　自由民主党　第160回国会　衆議院　憲法調査会　第1号　平成16年8月5日）

最後に他者の批判に言及する例を2つ挙げる。このうち，例 (7) の「よい表現で言えば感情的」は憲法前文が翻訳調とする声や改憲の提言者は現状の憲法を批判していることを婉曲的に述べている。また，例 (8) では，宮沢喜一は当初，日本国憲法の文体に違和感を持っていたが，徐々になじんでいったこ

とを述べている。さらに，好き嫌いの感情が憲法の文章に批判的か否かに影響することを示唆している。

(7)「憲法前文は翻訳調だという声があるんですね。それからまた，したがって，正しい日本語で書き直すべきだという，それに対する提言があるんです。どうも聞いておりますと，なかなかこれは気持ちがこもった発言であるがゆえに，よい表現で言えば感情的だというふうに申し上げてもよいと思います。」（土井たか子　社会民主党・市民連合　第162回国会　衆議院　憲法調査会　第4号　平成17年2月24日）

(8)「先ほど来，憲法の制定の経緯であるとか，あるいは前文についての好き嫌いも含めての意見の表明がありました。よく，現行憲法が翻訳調だというような批判のされ方があるんですけれども，私自身は実はそれを実感としてよくわからなかったんです。例えば，憲法二十三条は『学問の自由は，これを保障する。』五七五で書かれていますし，特に違和感を持って受けとめていなかったんですけれども，宮澤先生が，最初見たころはバタくさいなと感じたんだけれどもそのうちなじんでいったというような御発言をされていて，恐らく，制定当時，例えば手紙の書き方でも，候書きで教育を受けた方にとってみれば，バタくさいというのが非常に翻訳調だという言い方のあらわれなのかなと思ってお話を伺いました。」（山花郁夫　民主党・無所属クラブ　第161回国会　衆議院　憲法調査会　第4号　平成16年12月2日）

4.3 米国議会会議録

米国議会会議録中のtranslation（派生語を含む）を内容別に分類した結果を表6に示す。合衆国憲法の他の言語への翻訳を指す場合が最も多く，次いで，他国の憲法の英訳を指す場合が多くみられた。

前者の場合，どのケースも合衆国憲法の他言語訳およびその翻訳者を高く評価している点が特徴的であった。次の例はスペイン語訳の発行がもたらす恩恵について述べている。

(9) “I have not doubt that a House document in Spanish, especially the Constitution,

表 6. 米国議会会議録中の Constitution と共起する translation

種別	件数	内訳
合衆国憲法の他の言語への翻訳	10	子ども向け[13]（5），スペイン語訳（3），ハンガリー語訳（1），ギリシャ語訳（1）
他国の憲法の英訳	7	キューバ憲法（3），ニカラグア憲法（2），ブラジル憲法（1），ロシア憲法（1）
他の言語への翻訳以外の意味	5	
延べ数	22	-

注　「内訳」列の括弧内の数字は件数を示す。

can be an effective step towards this end. There is no doubt, as well, that a Spanish translation of the Constitution would also be useful to House Members in being able to provide copies for use in bilingual classes, Spanish language classes, and for use in naturalization classes, in addition to meeting the immediate need for a clearer understanding of the impeachment process. Above all, as true Representatives, we can use this effort to provide Spanish translations of our Constitution to increase communications with our Spanish speaking population.”（Patricia Schroeder, Colorado, House of Representatives; *Congressional Record*, vol. 120, part 21, Extensions of Remarks[14], 8/8/1974）

一方，後者の場合，該当国の憲法の英訳は，米国との貿易関係や外交関係上，その国の国家基盤を把握するために使われていた。議員が該当国の言語を理解できないためである。この場合，次に示す例のように議員は翻訳に対して否定的でも肯定的でもない。

（10）“Let me quote again—you heard it in Senator Borah’s speech, but it is of sufficient importance to be quoted again—article 106 of the Nicaraguan constitution in its entirety. I use the translation of the constitution published by the Government of the United States in 1919.”（Burton Kendall Wheeler; *Congressional Record*, vol. 68, part 2, Senate, 1/26/1927）

ただし，日本国憲法の議論とは違い，翻訳の自然さを議論するケースは本研究のデータ中にはなかったが，複数の翻訳を比較して，訳語の適切さを議論するケースはあった（例（11））。ここでは lack という語をその出現箇所を引用して（in the case…the office…）紹介している。

（11）“I have here the translation of the constitution of Nicaragua, which was sent by the State Department of the United States to the Committee on Foreign Relations. I find only one word in this translation different from the wording of the section as the Senator from Montana read it. That is the word “lack” instead of “default.” It reads:

In the case of the absolute or the temporary lack of the President of the Republic, the office—

And forth.”（Henrik Shipstead; *Congressional Record*, vol. 68, part 2, Senate, 1/26/1927）

このほかに，日本語では「翻訳（する）」に相当しない translation/translate の用例も複数あった。これらの用例では動詞 translate は日本語の「変える」「移す」などの動詞に相当する。例えば，“translate [the Constitution] into real life” がある。これらの用例は，Constitution が合衆国憲法を指し，into が動詞 translate に後続する点が共通していた。つまり，文構造は“V + O + into + 名”のように定式化できる。また，前置詞 into の目的語は変化・推移・行為の結果を示す名詞句がくることも共通している。一方，Constitution of Japan や Japan's Constitution など日本国憲法を意味する語は，今回のように翻訳の文脈に限定すると，米国議会会議録には用例が皆無であった。さらに，日本の国会で盛んに取り上げられた日本国憲法の英語版についても翻訳の文脈では言及がみられなかった。

5. まとめと結論

本論考では，現行憲法は翻訳の口調が残り，質が悪いという見解が憲法を実際に審議したあるいはしている議員にも共有されていたあるいはされているのか否かを検証するために，日本の憲法の文脈において翻訳が帝国議会あるいは

国会の場でどのように語られ，また，日本国憲法成立後と成立前でその語られ方に違いがあるかという研究設問を立て，議員の発言内容を分析し，その形式的･意味的な特徴を明らかにした。帝国議会および国会での議員の発言から「翻訳」を抽出し，分析した結果，帝国議会においては憲法の日本語原案，国会では現行憲法に対して翻訳の調子を持つとして否定的または批判的な見解が議員にみられた。加えて後者ではその否定や批判の強度が増し，また議員が自党を代表して批判するケースが新たに生じた。また，政党によって現行憲法に批判的な立場か，中立的な立場かが分かれる傾向があった。つまり，現行憲法が翻訳調であるという見方は議員の統一的な見方ではないといえる。対照事例として，米国議会において翻訳が憲法の文脈でどのように語られてきたかを同様に分析した結果，第一に，合衆国憲法の他言語訳に対しては肯定的であることをみた。この要因として次のようなことが指摘できる。合衆国憲法は模範憲法として，日本のみならず多くの国々の憲法の制定に影響を及ぼしてきた（Billias, 2009; Blaustein, 1987）。このように合衆国憲法が翻訳を通して世界に広まることは, 世界の大国としての米国の地位の確保に貢献すると推察される。つまり，米国にとって歓迎すべきことと考えられる。第二に，米国議会議員は他国の憲法の英語訳が文章として自然かどうかを問題にはしていなかったが，これはこの英語訳が他国の情報を得るための翻訳であって，必要な情報が得られれば翻訳の巧拙は求められないためと考えられる。さらに，日本語と英語の影響力の大きさも関係する。日本では，他国の憲法を日本語に翻訳することが多いが，自国の憲法を英語以外の言語に翻訳することは少ない。一方，米国は自国の憲法を他の言語に翻訳すること，すなわち多言語化することに力を入れる（National Constitution Center, n.d.）。他国の憲法は英語版がすでに存在することが多いからである。これらのことから，例えばリベリアのように（Blaustein, 1987: 21–22, 25）日本と同様に合衆国憲法から大きな影響を受けた国の憲法では翻訳についてどのような議論がなされているかを今後研究していくことが必要であろう。

　また，日本の現状憲法の国会での議論では，翻訳調を避けるべきものとしているが，全国紙（本論では『朝日新聞』と『読売新聞』）の記事の中で日本国憲法と翻訳調が共起するケースは1990年代になって初めて登場した。少なくとも1990年代までは全国紙を読む一般市民は日本国憲法が翻訳調であるか否かについて関心を払っていなかったと考えられる。これらの記事には，議員，特に自民党議員の現行憲法は翻訳調だという見解を取り上げたものが多い。一

般市民の見解を取り上げた記事は数少なく，それらの数少ない記事からは，一般市民が憲法の文体を受け入れたことがうかがえる（例えば，池澤（1996））。また，国会においても国会議員からではないが「現在の前文は，大変に格調高いものであり，国民の間にも定着しており，改正の必要は一切ないとの御主張もなされているところでございます」（橘幸信　衆議院法制局　第183回国会　衆議院　憲法審査会　第9号　平成25年5月16日）という指摘があった。広く日本の論壇に目を向けると法令文や技術文書の翻訳について「『翻訳調』として，しばしば問題になる『日本語』は，実は『科学』『社会』など無数の，外来の諸概念を実現するための新造語の創出とともに，それを可能にするための日本人の生み出した貴重な工夫であったと考えるべきであろう」（村上，2021: 96）といった前向きな見方もある。日本国憲法は翻訳調という紋切型の議論は再考の必要があると思われる。

謝辞

本稿の内容は，英語コーパス学会第48回大会（2022年10月，オンライン）における研究発表に基づいている。本稿の執筆に際し，査読委員3名の先生方より多くの貴重なご助言とご示唆を賜った。ここに深く感謝申し上げたい。本稿における不備はすべて筆者の責任である。

注

1. ただし，これは2023年現在では主な理由とはなっていない。例えば，『朝日新聞』が2022年に行った全国の有権者に対する世論調査によれば，改憲が必要とする理由として「国防の規定が不十分」（29%），「古くなった」（27%）が上位2位であり，「アメリカからの押し付け」は16%にとどまる（『朝日新聞』（2022.5.3））。
2. 「基」の読みの「もと」は縦書きの原典では漢字の右側に付されていたが，ここでは漢字の後ろの括弧内に表示している。
3. 本稿で使用する翻訳は，主として，議員が日本国憲法の論議に用いる場合の翻訳を指す。日本国憲法についてはさまざまな政治的立場やイデオロギー的立場から論じられているが，それが各議員の翻訳に対する見方に影響を及ぼすと考えられる。これは一般的に用いる翻訳とは異なる。しかし，前者の意味での翻訳か，一般的な意味の翻訳かのいずれを指すかは文脈で理解可能と思われるため，特に表記上区別しない。なお，鍵括弧は翻訳という語を本文中で特に強調したい場合に付す。
4. かつては「翻」の異体字「飜」から成る「飜訳」や「訳」の旧字体の「譯」から成る「翻譯」「翻譯」の表記がみられたが，本研究では「翻訳」とこれらを区別せ

ずに用いる。ただし，引用箇所は原典通りとする。

5. https://books.google.com/ngrams/
なお，検索するコーパスはAmerican English 2019，British English 2019，English 2019（https://books.google.com/ngrams/infoを参照）の3つを選択した。いずれもtranslationeseが圧倒的に多かった。

6. 二人は，地政学分野のイタリア語記事のタグ付きコーパスから一部を抽出した非翻訳記事と翻訳記事から成るテキストをサポートベクタマシン（SVM）に学習させた後，コーパスの残りの記事を自動で仕分ける実験を行った。その結果，正確度，精度，再現率がいずれも85%以上となったこと，さらに，同じ仕分けの作業を10名にさせたところ，正確度，精度，再現率の被験者のスコアの平均値よりもSVMのスコアの方が上回ったことを報告している。この結果の分析により，Baroni and Bernardini（2006）はSVMが非翻訳記事と翻訳記事とを仕分ける手掛かりとしていたのは，機能語と形態統語的要素の分布，とりわけ人称代名詞と副詞の分布であったことを見出した。

7. ANDはAND検索を示す。この検索式の仕様については，帝国議会会議録検索システムのヘルプ（使い方ガイド）（https://teikokugikai-i.ndl.go.jp/help.html）の「4. 検索機能の詳細」を参照。

8. ORはOR検索を示す。この検索式の仕様については，国会会議録検索システムのヘルプ（使い方ガイド）（https://kokkai.ndl.go.jp/help.html）の「4. 検索機能の詳細」を参照。

9. *は正規表現の数量子。直前の文字が0個以上であることを示す。

10. 以降，引用中の下線は筆者による。

11. 北の指摘は，（1）に抜粋している。一方，鈴木は憲法の文章を批判する際，翻訳という語を使っていない。「今回御提案の憲法前文…北君は,主として外國語臭い,別に外國語を參考にして書いたものではないと信じますが，外國語の臭ひがする，それも間違つた理解の上に立つて居ると云ふやうな點を御指摘になつたのでありまするが，之を讀みますると，洵に冗漫であり，切れるかと思へば續き，源氏物語の法律版を讀むが如き感がある…極端に申せば，泣くが如く，訴ふるが如く，嫋々として盡きざること縷の如しと言ひたい，一沫の哀調すら漂つて居るやうに感ずるのであります」（鈴木義男　日本社会党　第90回帝国議会　衆議院　本会議　第6号　昭和21年6月26日）

12. ここで段落末となっているため，句読点はない。

13. 子ども向けの「翻訳」は，日本語で一般的にいう外国語への翻訳とは異なるが，書名が*Constitution Translated for Kids*であり，著者のCathy Travisも自身のホームページで同書を“a simple, widely acclaimed, non ideological translation of the entire U.S. Constitution, side-by-side with the original 1787 text.”（https://www.travisbooks.com/children）と紹介し，自著をtranslationと捉えているため，本研究では他言語への翻訳と同様に扱った。一方，日本では，子ども向けに書かれた日本語の憲法の書物として,例えば『井上ひさしのこどもにつたえる日本国憲法』が刊行されている。

同書は書名には「翻訳」という語を用いていないものの，出版社による内容紹介には「平和憲法の精神を表している『前文』と『第九条』を，井上ひさしが子どもにも読める言葉に『翻訳』」（講談社 , 2006）とあり，「翻訳」が括弧付きで用いられている。

14. Extensions of Remarks とは米国議会の議場で行った声明を補完する賛辞，声明，その他情報を含む記録である。詳細は，GovInfo（https://www.govinfo.gov/help/crec）を参照。

参考文献

浅川哲也（2019）「国会会議録にみられる〈ら入れ言葉〉の使用実態について」『言語の研究』第 5 号：57–72.

Baroni, M., & Bernardini, S. (2006) “A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text.” *Literary and Linguistic Computing*, 21(3): 259–274.

Billias, G. (2009) *American Constitutionalism Heard round the World, 1776–1989: A Global Perspective*. New York, NY: New York University Press.

Blaustein, A. P. (1987) “Our Most Important Export: The Influence of the United States Constitution Abroad.” *Connecticut Journal of International Law*, 3(1): 15–30.

-ese, suffix. (2022, June). *OED Online.* Retrieved August 15, 2022, from www.oed.com/view/Entry/64342

-ese. (n.d.). In *Merriam-Webster.com Dictionary*. Retrieved February 23, 2023, from https://www.merriam-webster.com/dictionary/-ese

「不安定な世界，憲法は　朝日新聞社世論調査」（2022.5.3）『朝日新聞』朝刊，6 頁.

Furuno, Y. (2005) “Translationese in Japan.” In Hung, E. (ed.), *Translation and Cultural Change: Studies in History, Norms and Image-projection*. Amsterdam: John Benjamins, pp. 147–160.

伊原紀子（2000）「文学翻訳における異化・同化」『国際文化学』（神戸大学国際文化学会），03：105–116.

池澤夏樹（1996.7.29）「絶対安全の虚構　憲法を信じて進む文学的時代」『朝日新聞』夕刊，3 頁.

石黒圭（2007）『よくわかる文章表現の技術 V：文体編』東京：明治書院.

Kanoksilapatham, B. (2007) “Introduction to Move Analysis.” In Biber, D., Connor, U., & Upton, T. A. (eds.), *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins, pp. 23–41.

講談社（2006）「講談社 BOOK 倶楽部」Retrieved November 29, 2022 from https://bookclub.kodansha.co.jp/product?item=0000183309

松田謙次郎（2008）「国会会議録検索システム概論」松田謙次郎（編）『国会会議録を使った日本語研究』東京：ひつじ書房，pp. 1–32.

松田謙次郎（2012）「国会会議録をつかう」日比谷潤子（編）『はじめて学ぶ社会言語学―ことばのバリエーションを考える 14 章』京都：ミネルヴァ書房，pp. 54–79.

松村明 三省堂編修所（編）（2019）『大辞林』第 4 版　東京：三省堂.

Meldrum, Y. F. (2009a) *Contemporary Translationese in Japanese Popular Literature* [Doctoral dissertation, University of Alberta].

Meldrum, Y. F. (2009b) “Translationese in Japanese Literary Translation.” *TTR*, 22(1): 93–118. DOI: doi.org/10.7202/044783ar

Meldrum, Y. F. (2009c) “Translationese-specific Linguistic Characteristics: A Corpus-based Study of Contemporary Japanese Translationese.”『通訳翻訳への招待』3 号：111–131.

村上陽一郎（2021）『文化としての科学／技術』東京：岩波書店.

National Constitution Center (n.d.) *About us*. Retrieved November 29, 2022 from https://constitutioncenter.org/about

Newmark, P. (1991) *About Translation*. U.K.: Multilingual Matters.

大岡玲（2017）「『翻訳』というアイデンティティ」『日本語学』第 36 巻，第 12 号：8–17.

新村出（編）（2018）『広辞苑』第 7 版　東京：岩波書店.

小学館辞典編集部（編）（2003）『使い方の分かる類語例解辞典』新装版　東京：小学館.

Swales, J. (1990) *Genre Analysis: English for Academic and Research Settings*. Cambridge: Cambridge University Press.

Tirkkonen-Condit, S. (2002) “Translationese: A Myth or an Empirical Fact? A Study into the Linguistic Identifiability of Translated Language.” *Target*, 14(2): 207–220.

Translatese. (2022, June) *OED Online*. Retrieved August 15, 2022, from https://www.oed.com/view/Entry/204842

Translationese. (2022, June) *OED Online*. Retrieved August 15, 2022, from https://www.oed.com/view/Entry/204846

Translatorese. (2022, June) *OED Online*. Retrieved August 15, 2022, from https://www.oed.com/view/Entry/204850

Translatorese. (n.d.) Translatorese. In *Merriam-Webster.com Dictionary*. Retrieved February 23, 2023, from https://www.merriam-webster.com/dictionary/translatorese

梅林博人（2000）「近代小説にみる接続詞『そして』　翻訳調といわれる『A そして B』をめぐって」『国文学　解釈と鑑賞』65–07：36–42.

山口昌也（2017）「国会会議録における言語表現の時間的変化の予備的分析」『言語資源活用ワークショップ発表論文集』2：304–312.

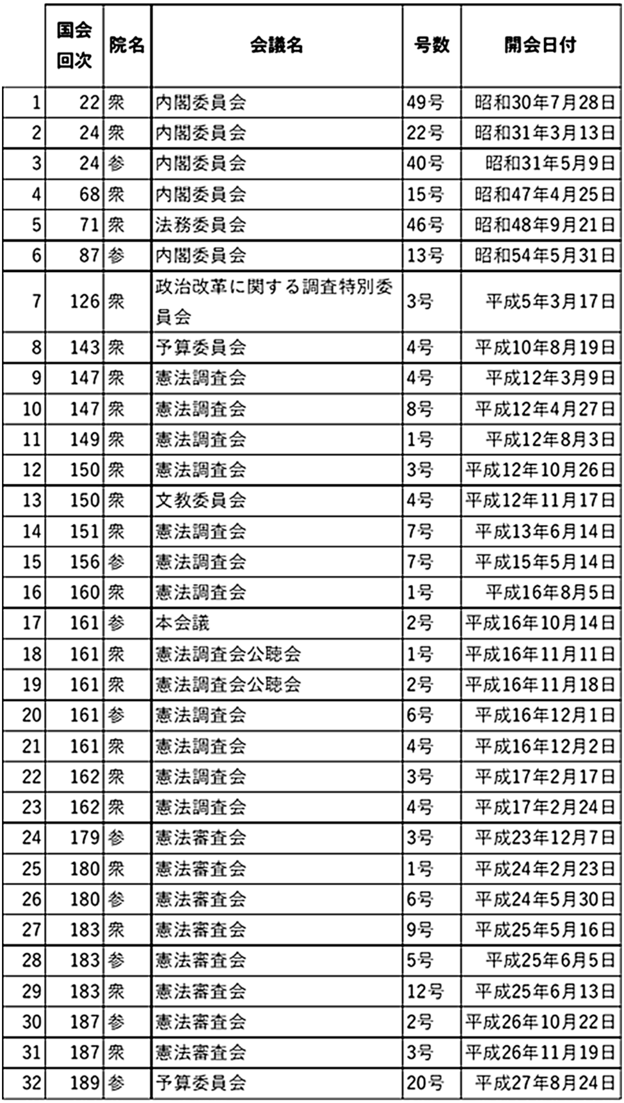
（島津美和子　立教大学）

付録

表 1. 使用した帝国議会会議録

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

表 2. 使用した国会会議録



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

「研究ノート」

Measuring the Effects of Frequency and Dispersion on Key Expression Analysis: Methodological Recommendations

Yuichiro KOBAYASHI

Abstract

The automatic identification of key expressions is among the most crucial tasks in corpus linguistics. The use of random forest (RF) for analyzing key expressions has been increasing lately owing to its ability to efficiently compute the keyness scores of many linguistic features. However, only little is known of the types of predictor variables that RF assigns high scores to. It is significantly risky to rely on a single method without understanding its pros and cons. Thus, this study was conducted to distinguish between the two most employed machine-learning algorithms for variable selection – RF and least absolute shrinkage and selection operator (LASSO) – as the tools for extracting key expressions from corpora. Specifically, employing the corpus of *PLOS ONE* research articles, both statistical methods were compared regarding the effects of frequency and dispersion. The results indicate that RF selects the high-frequency variables in a large number of texts, while LASSO selects the low-frequency variables in a small number of texts. Since both methods have their pros and cons, the purpose of the research will determine the method to be adopted.

1. Introduction

Corpus analysis generally involves the comparison of multiple text groups, as well as the extraction of key expressions, such as keywords and key *n*-grams, that characterize each group (Scott & Tribble, 2006). Specifically, corpus-based stylistic studies differentiate the language use of a particular writer or fictional character from the linguistic *norms* represented by general or reference corpora (Stubbs, 2005). By comparing learners at different proficiency levels, learner corpus studies also reveal their characteristics at each level (Pérez-Paredes & Díez-Bedmar, 2019). Furthermore, studies on English for specific purposes have produced a list of the key expressions in a particular academic field by analyzing the corpus of research articles on multiple academic disciplines (Asano, 2018). Regardless of the differences in the analyzed registers, these studies were all aimed at identifying the key expressions that are more salient in a group of texts than in others.

Various statistical methods have been employed to identify such key expressions. Conventionally, most analyses of key expressions have relied on the log-likelihood ratio and chi-square tests that are implemented in most corpus analysis tools. However, these tests cannot account for the variance within a group since the individual texts within a group are aggregated at the group level. For example, when distinguishing between the styles of two writers, the characteristics of their individual texts are generally ignored, and the writers are directly compared. This represents a limitation that several researchers have attempted to overcome via mean and median tests (Paquot & Bestgen, 2009) or dispersion measures (Egbert & Biber, 2019; Gries, 2021) for keyword extraction.

A recent trend in keyness analysis is the application of machine-learning methods in the detection of linguistic features that can predict text groups or linguistic choices. Particularly, the utilization of random forest (RF) models (Breiman, 2001) has been increasing in several areas of corpus linguistics, such as grammatical studies (Deshors, 2019; Deshors & Gries, 2016; Hundt et al., 2020; Paquot et al., 2019), sociopragmatics (Funke & Bernaisch, 2022), stylometry (Suzuki & Hosoya, 2014; Tabata, 2014), and learner corpus research (Kobayashi & Abe, 2016; Kobayashi et al., 2022; Tono, 2013). RF can efficiently analyze thousands of linguistic features and compute variable importance scores for each feature, representing the magnitude of the differences between the frequencies of the compared text groups or linguistic choices.

However, it is dangerous to depend on a single method without understanding its advantages and disadvantages. RF models are not always stable or robust, and even small changes in the dataset can significantly change the results of the analysis (Gries, 2020). In addition, the changes in the method can also significantly impact the results of the data analysis (Silberzahn et al., 2018). In the field of corpus linguistics, it is well-known that the choice of collocation measures greatly affects the results of the collocation analysis. Generally, collocations with high *t*-scores exhibit high-frequency pairs, whereas those with high mutual-information scores include low-frequency words (McEnery et al., 2006). This knowledge of collocation statistics reveals the risks of relying on a single method. Thus, it is very crucial to understand the differences in key expression analyses that are performed with different statistical methods.

In machine learning, key expression analysis is considered an application of variable selections that identifies crucial predictor variables for high-accuracy data classification. The two most widely employed algorithms for variable selection are RF and the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996). Both statistical methods have achieved high success in data analysis competitions, such as Kaggle, because they generally offer good predictive performance and low overfitting (Banerjee, n.d.). Owing to their abilities to analyze many variables efficiently, both methods also exhibit great potential in text analysis for measuring the keyness of all words or *n*-grams in the corpus.

2. RF and LASSO

RF consists of hundreds or thousands of decision trees; each tree is built via the random samplings of the observations from the dataset and predictor variables. The variables are not completely considered by every tree, which renders the trees decorrelated and less prone to overfitting. The RF model performs a final prediction by synthesizing the results of each tree using the majority vote. In this model, the mean decrease in the Gini coefficient is employed as a measure of the contribution of each variable to the prediction: the higher the value of the mean decrease in the Gini coefficient, the higher the significance (keyness) of the variable in the model.

LASSO is a modeling method that simultaneously performs variable selection and model building. Technically, this method employs L1 regularization, which adds a penalty that is equivalent to the absolute value of the magnitude of the coefficients (Hastie et al., 2009). By adding the penalty, LASSO can avoid overfitting and remove the insignificant predictor variables from the dataset. The resulting reduced number of variables enhances the prediction accuracy and interpretability of the model. In the LASSO model, the partial regression coefficients of each variable can be used as the keyness scores for the key expression analysis.

3. Purpose of the Study

This study was conducted to compare two statistical methods, RF and LASSO, as tools for extracting key expressions from corpora. Particularly, both methods were compared in terms of frequency and dispersion. In this study, the term *key expressions* is defined as linguistic features that are instrumental in the statistical discrimination of two or more text groups. The term *frequency* indicates the number of times an expression is used in the text, and *dispersion* refers to the number of texts a given expression appeared in. The following research question was explored here: How do RF and LASSO differ regarding the effects of frequency and dispersion? The answer to this question can offer methodological recommendations with which corpus linguists can select appropriate keyness measures for their research.

4. Corpus

In this study, a corpus of 1,000 *PLOS ONE* research articles was employed on machine learning. The corpus was compiled with AntCorGen (version 1.2.0) (Anthony, 2022), which was the latest version at the time of conducting the study. To extract the key expressions from the corpus, the introduction (INT) and results and discussion (RAD) sections of the 1,000 articles were compared. Table 1 reveals the numbers of texts and words in each section analyzed in this study.

Table 1. Numbers of texts and words per section

	Number of texts	Number of words
Introduction (INT)	1,000	1,403,982
Results and discussion (RAD)	1,000	4,486,595
Total	2,000	5,890,577

5. Data Analysis

In this study, the relative frequencies (per 100 words) of 65 of 67 Biber's (1988) linguistic features were counted using the Multidimensional Analysis Tagger (Nini, 2019). The values of the remaining two variables – type/token ratio (TTR) and mean word length (AWL) – were also calculated by the tool. Thereafter, these 67 values were employed to compare the INT and RAD sections. All statistical analyses were conducted using R (version 4.2.0), a free software environment for statistical computation and graphics (R Core Team, 2021). The *randomForest* and *glmnet* packages were used to perform the RF and LASSO analyses, respectively. Further, rank sum tests, correlation analysis, and linear regression analysis were conducted to complement the results of RF and LASSO.

6. Results and Discussion

6.1 Key Expression Analysis via RF

This study began with the key expression analysis using RF. The RF model employed Biber's 67 linguistic features as the predictor variables and utilized two text types (INT and RAD) as the response variables; the results of the classification are presented in Table 2, in which the columns and rows of the matrix represent the text types predicted by the model and the actual text types, respectively. Thus, the column "Accuracy" indicates the agreement rates between the predicted and actual text types. In this classification model, high accuracy indicates significant differences in the frequency patterns of the linguistic features between the INT and RAD sections.

Table 2. Accuracy rates of the RF models of the two text types

	INT (predicted)	RAD (predicted)	Accuracy
INT (actual)	956	44	95.6%
RAD (actual)	26	974	97.4%

Note. Overall accuracy rate evaluated via out-of-bag simulation was 96.5%.

RF can be used to compute the keyness scores, also known as the variable importance scores, for measuring the impact of each predictor variable on the alternation, given all the other predictors. The keyness scores demonstrated that the top-five linguistic features that distinguished the text types were (a) PEAS (perfect aspect), (b) TO (infinitives), (c) VPRT (present tense), (d) AWL (mean word length), and (e) VBD (past tense). The Wilcoxon rank sum tests with continuity correction also indicated that significant differences existed between the INT and RAD sections of all five variables ($p < 0.001$). One of the easiest strategies for identifying the linguistic features that characterize each section is to draw their box plots. Figure 1 shows the box plots of the frequency distributions of the top-five linguistic features. The plots revealed that PEAS, TO, VPRT, and AWL were characteristic of INT, while VBD was characteristic of RAD.

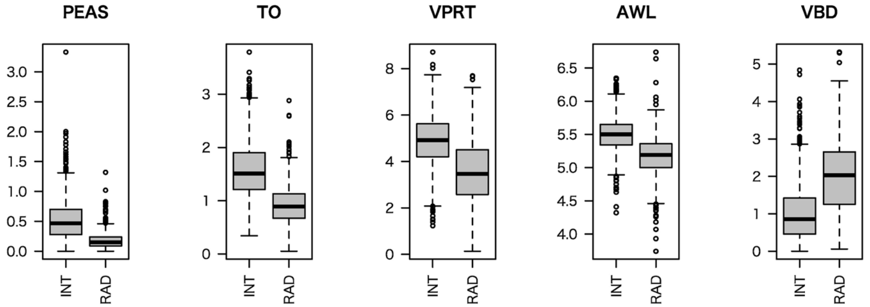


Figure 1. Frequency distributions of the top-five linguistic features in the RF model

The key expression analysis using RF is generally performed via the aforementioned procedure. However, the types of predictor variables that RF assigns high scores to are barely known. Therefore, this study was aimed at elucidating the relationships among the keyness scores, frequencies (or the magnitude of the values), and dispersion values of each predictor variable. Figure 2 visualizes the relationship between the keyness scores and relative frequency, as well as between the keyness scores and dispersion in the RF model. The horizontal and vertical axes of the figures indicate the logarithm of the keyness scores and frequency and dispersion values of the 67 linguistic features, respectively. The straight lines, representing the results of the linear regression analysis, indicated that RF tends to assign high scores to frequently used very dispersed variables. Pearson's product-moment correlation coefficients also revealed high, positive correlations between the keyness scores and frequency values (r = 0.57) and between the keyness scores and dispersion values (r = 0.65). Put differently, RF tends not to extract key expressions that are prominently utilized in a particular text type, albeit at low frequency or only in a few texts. This tendency is generally observed in other decision-tree-based models including gradient boosting (Friedman, 2001). Moreover, RF generates keyness scores for all predictor variables, although no theoretical threshold can be used to discriminate between the relevant and irrelevant variables. Thus, the threshold setting generally tends to become arbitrary. Despite these shortcomings, this method can be useful for studies of academic English that identify frequent words and phrases in each section of academic articles. The method also can be instrumental in contrasting the language use of learners at different proficiency levels.

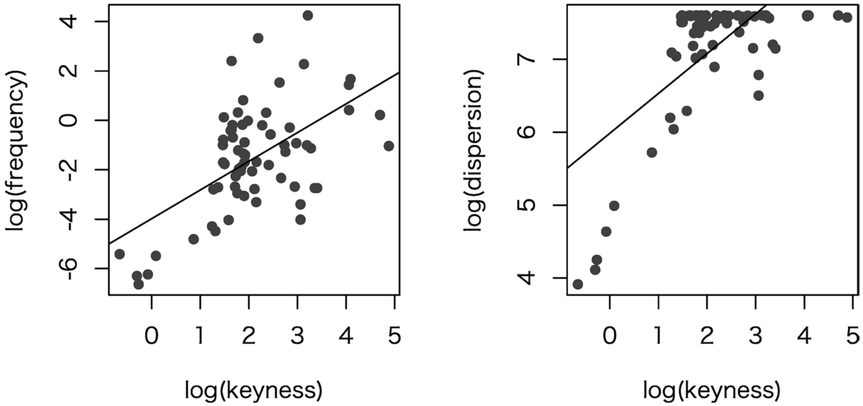


Figure 2. Relationships among the keyness, relative frequency, and dispersion in the RF model

6.2 Key Expression Analysis via LASSO

Owing to the threshold-setting issue in RF, LASSO was considered as the tool for key expression analysis, and Table 3 presents the classification results of the LASSO model using 67 linguistic features and two text types as the predictors and responses, respectively. Since the overall accuracy rate was 95.6%, modeling with LASSO was as highly reliable as that with RF (96.5%).

Employing the variable selection process in the LASSO model, 25 of the 67 predictor variables were statistically selected. The keyness scores (i.e., the absolute

Table 3. Accuracy rates of the LASSO model for the two text types

	INT (predicted)	RAD (predicted)	Accuracy
INT (actual)	952	48	95.2%
RAD (actual)	40	960	96.0%

Note. Overall accuracy rate evaluated via cross validation was 95.6%.

values of the coefficients) indicated that the top-five linguistic features that can distinguish text types were (a) SMP (*seem* and *appear*), (b) THAC (*that* adjective complements), (c) DPAR (discourse particles), (d) TOBJ (*that* relative clauses on object position), and (e) STPR (stranded prepositions). These five linguistic features do not overlap with a single feature in the top-five features of the RF model. Figure 3 shows the frequency distributions of the top-five variables in the LASSO model. The values on the vertical axes indicate that the low-frequency variables were assigned high scores. Additionally, the medians of one or both text types for the five variables were zero. However, the Wilcoxon rank sum tests with continuity correction detected significant differences between the text types for all five variables ($p < 0.001$).

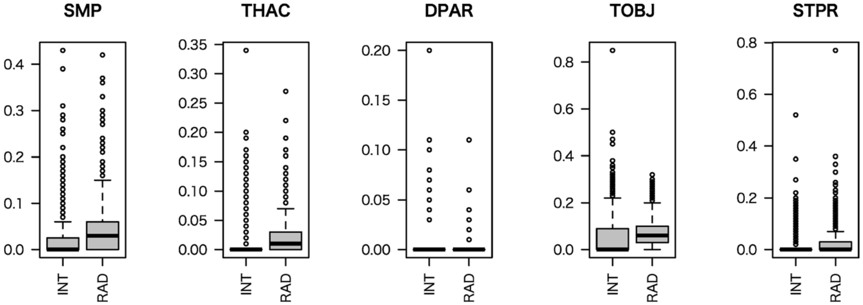


Figure 3. Frequency distributions of the top-five linguistic features in the LASSO model

Figure 4 shows the relationship between the keyness and relative frequency and between the keyness and dispersion in the LASSO model. Pearson's product-moment correlation coefficients also indicated the high, negative correlations between the keyness scores and frequency values ($r = -0.66$) and between them and the dispersion values ($r = -0.65$). The results indicate that, compared with RF, LASSO tends to assign high scores to variables that are used at low frequencies in a small number of texts. From a linguistics standpoint, while the RF model highlighted the difference in frequency of features such as tense and aspect markers (PEAS, VPRT and VBD), the LASSO model emphasized the difference in features such as stance markers (SMP) and discourse markers (DPAR). Owing to its nature, in stylometry, LASSO can identify the specific linguistic features that an author uses in comparison to other authors.

Furthermore, in the studies of English for specific purposes, this method can detect idiosyncratic words and phrases in a particular academic or professional field.

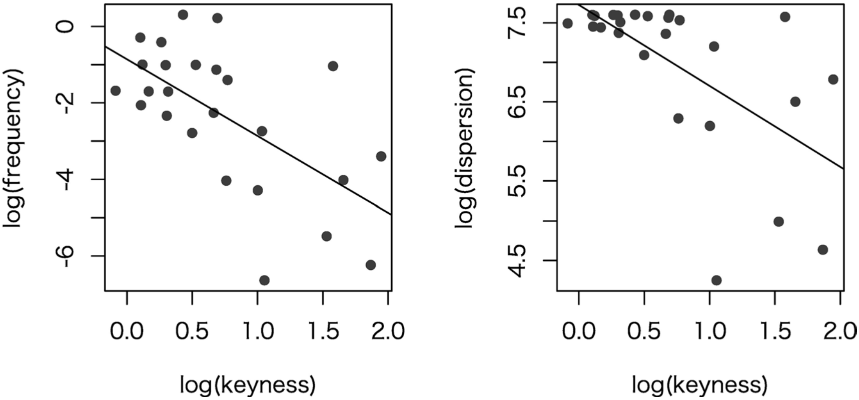


Figure 4. Relationships among the keyness scores, relative frequency, and dispersion in the LASSO model

7. Conclusion

This study was conducted to compare RF and LASSO as tools for analyzing key expressions from frequency and dispersion viewpoints. The results indicate that the keyness scores of RF and LASSO are positively and negatively related to the frequency and dispersion values, respectively. In many cases, RF models with a tendency to select high-frequency variables may be easier to interpret than LASSO models. However, the high-frequency features might be predictable without performing a statistical analysis. Additionally, as previously mentioned, the models cannot explicitly distinguish relevant variables from irrelevant ones. Conversely, the LASSO models with a tendency to select low-frequency variables could offer new insights that transcend the analyst's predictions. Furthermore, the LASSO models can remove irrelevant variables from the analysis. Nevertheless, the models only select one of the variables when there is a pair of highly correlated variables (Freijeiro-González et al., 2022). Therefore, if no substantial difference exists between the classification accuracies of both algorithms, the choice between them depends on the purposes of the intended studies. To deeply understand the differences between RF and LASSO, the methods must be compared via various types of datasets and measures other than frequency and dispersion. Additionally, it is interesting to consider several improved versions of the algorithms, such as Boruta (Kursa & Rudnicki, 2010) and elastic net (Zou & Hastie, 2005), as tools for comparison. Boruta can clearly distinguish between relevant and irrelevant variables by embedding statistical tests in the RF model. Elastic net can assign the same keyness scores to highly correlated variables by combining LASSO and ridge regression models. Moreover, other statistical techniques than the RF- and LASSO-related methods should be applied to seek a better set of key expressions. Similar to the findings on the collocation statistics, the knowledge of the appropriate use of multiple keyness measures can increase the validity of statistical analysis in corpus linguistics.

Acknowledgements

This research has been partly funded by Grants-in-Aid for Scientific Research (Grant Number: 21K00660). I would like to express my gratitude to the two anonymous reviewers for their helpful comments and suggestions.

References

Anthony, L. (2022). AntCorGen (Version 1.2.0). Waseda University. https://www.laurenceanthony.net/software

Asano, M. (2018). Construction of medical research article corpora with AntCorGen: Pedagogical implications. *English Corpus Studies, 25,* 101–115.

Banerjee, P. (n.d.). Comprehensive guide on feature selection. Kaggle. https://www.kaggle.com/code/prashant111/comprehensive-guide-on-feature-selection

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Deshors, S. C. (2019). English as a lingua franca: A random forests approach to particle placement in multi-speaker interactions. *International Journal of Applied Linguistics, 30*(2), 214–231.

Deshors, S. C., & Gries, S. Th. (2016). Profiling verb complementation constructions across New Englishes: A two-step random forests analysis to *ing* vs. *to* complements. *International Journal of Corpus Linguistics, 21*(2), 192–218.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77–104.

Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review, 90*(1), 118–145.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Funke, N., & Bernaisch, T. (2022). Intensifying and downtoning in South Asian Englishes: Empirical perspectives. *English World-Wide, 43*(1), 33–65.

Gries, S. Th. (2020). On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory, 16*(3), 617–647.

Gries, S. Th. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics, 9*(2), 1–33.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hundt, M., Rautionaho, P., & Strobl, C. (2020). Progressive or simple? A corpus-based study of aspect in World Englishes. *Corpora, 15*(1), 77–106.

Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics, 20*(1), 55–73.

Kobayashi, Y., Abe, M., & Kondo, Y. (2022). Exploring L2 spoken developmental measures: Which linguistic features can predict the number of words? *English Corpus Studies, 29,* 1–18.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13.

McEnery, A., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* Routledge.

Nini, A. (2019). The multi-dimensional analysis tagger. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 67–94). Bloomsbury Academic.

Paquot, M., & Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A. Jucker, D. Schreier, & M. Hundt (Eds.), *Corpora: Pragmatics and discourse* (pp. 247–269). Rodopi.

Paquot, M., Grafmiller, J., & Szmrecsanyi, B. (2019). Particle placement alternation in EFL learner vs. L1 speech: Assessing the similarity of probabilistic grammars. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (Eds.), *Widening the scope of learner corpus research: Selected papers from the Fourth Learner Corpus Research Conference* (pp. 71–92). Presses universitaires de Louvain.

Pérez-Paredes, P., & Díez-Bedmar, M. B. (2019). Researching learner language through POS keyword and syntactic complexity analyses. In S. Götz & J. Mukherjee (Eds.), *Learner corpora and language teaching* (pp. 101–128). John Benjamins.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.r-project.org/

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language* *education.* John Benjamins.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1,* 337–356.

Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature, 14*(1), 5–24.

Suzuki, T., & Hosoya, M. (2014). Computational stylistic analysis of popular songs of Japanese female singer-songwriters. *Digital Humanities Quarterly, 8*(1). http://www.digitalhumanities.org/dhq/vol/8/1/000170/000170.html

Tabata, T. (2014). Stylometry of Dickens's language: An experiment with random forests. In P. L. Arthur & K. Bode (Eds.), *Advancing digital humanities: Research, methods, theories* (pp. 28–53). Palgrave Macmillan.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58,* 267–288.

Tono, Y. (2013). Criterial feature extraction using parallel learner corpora and machine learning. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 169–204). John Benjamins.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B, 67*(2), 301–320.

Appendix: The 67 linguistic features from Biber (1988)

A. Tense and aspect markers

1. past tense (VBD), 2. perfect aspect (PEAS), 3. present tense (VPRT)

B. Place and time adverbials

4. place adverbials (PLACE), 5. time adverbials (TIME)

C. Pronouns and pro-verbs

6. first person pronouns (FPP1), 7. second person pronouns (SPP2), 8. third person personal pronouns (excluding *it*) (TPP3), 9. pronoun *it* (PIT), 10. demonstrative pronouns (DEMP), 11. indefinite pronouns (INPR), 12. pro-verb *do* (PROD)

D. Questions

13. direct WH-questions (WHQU)

E. Nominal forms

14. nominalizations (ending in *-tion, -ment, -ness, -ity)* (NOMZ), 15. gerunds (GER), 16. total other nouns (NN)

F. Passives

17. agentless passives (PASS), 18. *by*-passives (BYPA)

G. Stative forms

19. *be* as main verb (BEMA), 20. existential *there* (EX)

H. Subordination features

21. *that* verb complements (THVC), 22. *that* adjective complements (THAC), 23. WH clauses (WHCL), 24. infinitives (*to*-clause) (TO), 25. present participial clauses (PRESP), 26. past participial clauses (PASTP), 27. past participial WHIZ deletion relatives (WZPAST), 28. present participial WHIZ deletion relatives (WZPRES), 29. *that* relative clauses on subject position (TSUB), 30. *that* relative clauses on object position (TOBJ), 31. WH relatives on subject position (WHSUB), 32. WH relatives on object position (WHOBJ), 33. pied-piping relative clauses (PIRE), 34. sentence relatives (SERE), 35. causative adverbial subordinators (*because*) (CAUS), 36. concessive adverbial subordinators (*although, though*) (CONC), 37. conditional adverbial subordinators (*if, unless*) (COND), 38. other adverbial subordinators (OSUB)

I. Prepositional phrases, adjectives, and adverbs

39. total prepositional phrases (PIN), 40. attributive adjectives (JJ), 41. predicative adjectives (PRED), 42. total adverbs (RB)

J. Lexical specificity

43. type/token ratio (TTR), 44. mean word length (AWL)

K. Lexical classes

45. conjuncts (CONJ), 46. downtoners (DWNT), 47. hedges (HDG), 48. amplifiers (AMP), 49. emphatics (EMPH), 50. discourse particles (DPAR), 51. demonstratives (DEMO)

L. Modals

52. possibility modals (POMD), 53. necessity modals (NEMD), 54. predictive modals (PRMD)

M. Specialized verb classes

55. public verbs (PUBV), 56. private verbs (PRIV), 57. suasive verbs (SUAV), 58. *seem* and *appear* (SMP)

N. Reduced forms and dispreferred structures

59. contractions (CONT), 60. subordinator *that* deletion (THATD), 61. stranded prepositions (STPR), 62. split infinitives (SPIN), 63. split auxiliaries (SPAU)

O. Coordination

64. phrasal coordination (PHC), 65. independent clause coordination (ANDC)

P. Negation

66. syntactic negation (SYNE), 67. analytic negation (XX0)

（小林雄一郎　日本大学）

「研究ノート」

Development and Revision of DDL Tools for Secondary School Students: What We Can Do to Nurture Autonomous Corpus Users?

Chikako NISHIGAKI and Shiro AKASEGAWA

Abstract

While the effectiveness of using data-driven learning (DDL) has been well established, it has been mainly used with university level learners. In an effort to harness these advantages with younger learners, two specialized corpora and tools (eDDL and hDDL) have been developed and implemented over a number of years, called the DDL Project. In this paper, we provide an overview of our approach in developing and modifying hDDL, and describe the four stages undertaken to help students transition from using pedagogically modified corpora with teacher support to being able to use authentic corpora independently. Our goal has been not only to be able to implement DDL in EFL classes at the secondary school level to improve learners' knowledge of grammar, but to support learners in developing the skills to understand and navigate corpora themselves in order to go beyond prescribed grammatical targets to use corpora to address any language curiosities they may have. To further this goal, both eDDL and hDDL are available without cost or registration, so learners and teachers can access these online at any time.

1. Introduction

Data-driven learning (DDL) is a corpus-based foreign language teaching method said to be "the most promising applications of corpus linguistics" (Wicher, 2020, p. 31). With the support of a Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research, we began a research project (called the DDL Project) for secondary school learners in 2013 and expanded this to elementary schools in 2016. This DDL Project had three initial goals: creating pedagogical corpora suitable for elementary and secondary school learners, developing and revising appropriate corresponding corpus search tools and implementing DDL using these tools in school settings (Nishigaki et al., 2020, 2022). The developed tools are eDDL for elementary school (https://e.ddl-study.org/) and hDDL for secondary school or high school (https://h.ddl-study.org/). This current article describes our subsequent focus: the development and revision of the hDDL tool, which is intended to achieve the long-term goal of fostering future independent corpus users. Section 2 of this paper provides the background of our DDL project. Section 3 describes the development of the hDDL tool, and section 4 discusses the latest revisions. We provide an analysis of the revisions in section 5 and a conclusion in section 6.

2. Background

DDL is a means to apply corpus linguistics methods to language teaching. A learner searches a corpus to discover patterns and rules of a language, and constructs and learns knowledge about the language in his or her own way, just as a linguist would. In DDL classrooms, the learner's mission is to discover language facts and to modify or replace previous knowledge with new knowledge on his or her own, while the teacher's role is to help the learner discover facts and build new knowledge. This characteristic of DDL overlaps with the constructivist view of learning and with the Japanese Ministry of Education's (MEXT) "course of study" (Suzuki, 2021).

DDL, first proposed by Johns (1991), can be applied to a wide range of areas in foreign language learning; for example, to learning vocabulary (Tsai, 2019), collocations (Saeedakhtar et al., 2020), grammar (Lin, 2021; Mull & Conrad, 2013; Vyatkina, 2013), and writing (Elmansi et al., 2021). It can be used in various ways; for example, the language data can be computer-based, wherein users explore a corpus on a screen, or paper-based, wherein text data is pre-selected and printed for users (Gabrielatos, 2005); and the corpora can be parallel (comparing two languages) or monolingual. The effectiveness of DDL has been well established by a number of studies, including three meta-analyses (Boulton & Cobb, 2017; Lee et al., 2019; Mizumoto & Chujo, 2015).

Although DDL has many applications and has been shown to be effective, it is also the case that it is used primarily with university students or intermediate and advanced level learners. The use of DDL at the pre-tertiary level is not widespread (Crosthwaite, 2020). One reason for this is that there are no corpora suitable for pre-tertiary groups (Perez-Paredes, 2020). Furthermore, the use of simplified corpora has been criticized, for example, by Sinclair (1991, p. 6), who believed the authenticity of corpora is paramount: “One does not study all of botany by making artificial flowers.” Gabrielatos (2005) rebutted this thinking as problematic “corpus worship” (p. 20). We believe that if DDL is to be accessible to and effective with younger learners, appropriate corpora need to be created and used, at least initially, so learners can understand what corpora is and how to navigate it. In other words, there needs to be a bridge from using a simplified corpus to eventually using authentic corpora.

A second problem that hinders DDL implementation is the difficulty students have in using DDL search software. DDL requires a corpus, appropriate software, the ability to use that software, and a user-friendly interface (Timmis, 2015). However, secondary school learners and teachers new to DDL are not accustomed to using currently available corpus search software. For these reasons, the search software needs to be as simple and user-friendly as possible.

To address these issues, Chujo et al. (2015) developed the Sentence Corpus of Remedial English (SCoRE), which is a simplified DDL tool that was created for beginner level EFL university students. The English sentences in SCoRE were created from a specially made source corpus (a collection of beginner level English) and each concordance line is an independent and complete sentence. SCoRE was created as a needs-driven, classroom-ready resource, which is easy to use for both learners and teachers. Learners can search for English sentences containing a target grammatical item within a simple interface.

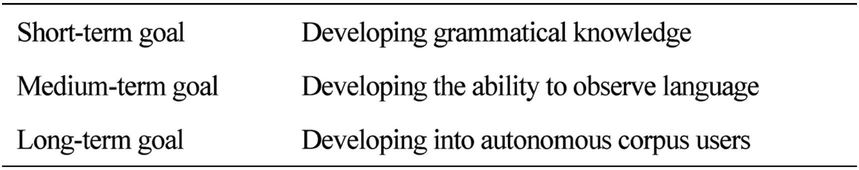
3. Development of Secondary School DDL Tools

Based on the results of second language acquisition (SLA) research, the Japanese course of study encourage English classes to be taught with a meaning-oriented approach (MEXT, 2017). However, while meaning-oriented interaction may result in better communication ability, communication without attention to language forms does not necessarily improve linguistic accuracy (Loewen, 2020). Therefore, English teachers need to consciously direct their students’ attention to grammatical items in the classroom while keeping the emphasis on communication. Since DDL explicitly and inductively teaches grammatical knowledge, we considered DDL to be an effective teaching method for grammar. As a result, we incorporated DDL into pre-tertiary level communicative English classes to teach grammar.

3.1 Setting DDL Goals

When conducting pedagogical activities, it is essential to set educational goals to define the abilities to be developed in learners. Our initial goal for the DDL Project was to create a level-appropriate pedagogical corpus, develop a user-friendly corpus search tool, and implement these to measure effectiveness and understand how these could be modified to be more effective. In this next stage of the DDL Project, our focus has been to continue to evaluate and modify these tools, but also to explore how to nurture secondary school learners in using DDL such that although they begin with a simplified pedagogical corpus and teacher support, they eventually are able to use authentic corpora independently. These goals are shown in Table 1, with timeframes divided into short-term, medium-term, and long-term goals according to the time it may take to achieve them.

Table 1. Pedagogical Goals for the DDL Project



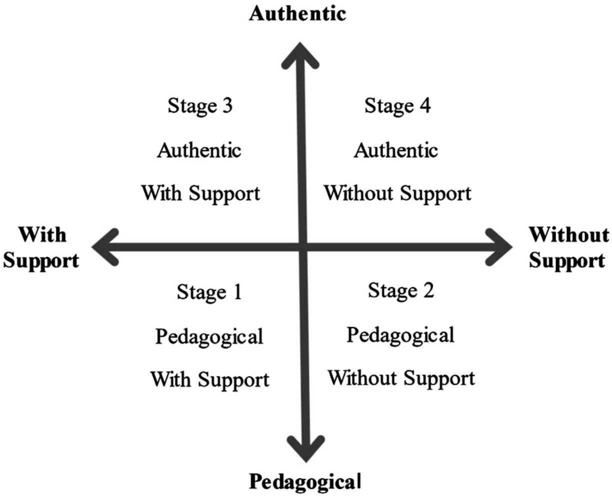
Short-term goal	Developing grammatical knowledge
Medium-term goal	Developing the ability to observe language
Long-term goal	Developing into autonomous corpus users

The short-term goal is to understand and acquire knowledge of the grammatical rules of the learning target, to be achieved in one class or one unit in the textbook (see previous studies Kakiba et al., 2021; Nishigaki et al., 2021; Nishigaki & Kakiba, 2023). The medium-term goal is to develop students' ability to observe English by asking themselves at what and how they look at the language when a question comes into their minds, and is achieved through the continuous implementation of DDL in class across a semester or a year (Nishigaki et al., 2018). The long-term goal is to develop autonomous corpus users. Since a variety of corpora can be accessed online, this is the ability to identify a corpus and obtain the information they need to solve their language-related question. Developing autonomous corpus use would be a lifelong language learning skill.

3.2 Setting the DDL Stage

One of the strengths of DDL is that it provides the learner with large amounts of authentic data, and learners can see how language is actually used. A corpus based on authentic data is effective because the language is more closely related to learners' needs and interests, and can enhance students' motivation, cover current issues, and support a more creative approach to teaching, among other benefits (Beresova, 2015). However, this strength of DDL—its authenticity in raw form—can be a disadvantage for beginning and elementary level learners because authentic corpora usually contain complex grammar, high level vocabulary, and multiple grammatical exceptions and are therefore challenging for lower level language learners. To create a bridge to authentic corpus use, pedagogical intervention is necessary. There are a variety of possible pedagogical interventions; for example, one is for the instructor to support the learner by selecting and presenting a corpus that is level-appropriate. For learners who are unfamiliar with the particular language format of concordance lines or who do not yet have sufficient linguistic knowledge of the target language, it is helpful to give them clues that elicit discovery. With these interventions in mind and in keeping with the goal of developing independent corpus users (see Table 1), we divided the learning progression into four stages (see Figure 1). The vertical axis is the type of corpus: authentic or pedagogical. The horizontal axis is the level of pedagogical intervention provided by the teacher (e.g., whether the learner is given a search guide).

In Stage 1, learners use the pedagogical corpus with teacher assistance. In Stage 2, learners use a pedagogical corpus without teacher assistance. In Stage 3, learners use an authentic corpus with teacher assistance. In Stage 4, learners use an authentic corpus without teacher assistance. We are applying these four stages to the context of English education in Japan as follows. Students begin to learn grammar in explicit ways in the 7th grade and learn English each year up to the 12th grade using government-approved textbooks. The 7th to 12th grades (secondary school) are divided into two categories in Japan: the 7th to 9th grades are junior high school, and the 10th to 12th grades are high school. Therefore, Stage 1 and Stage 2, which use pedagogical corpora, can be applied to 7th to 9th grades and 10th to 12th grades, respectively. The A1 level of the Common



The chart was revised from Nishigaki et al. (2015).

Figure 1. DDL Stages

European Framework of Reference for Languages (CEFR) is the MEXT goal for students by the time they graduate from the 9th grade, and the A2 level by the time they graduate from the 12th grade. Thus, as a rough guide, Stage 1 corresponds to the CEFR A1 level, and Stage 2, the CEFR A2 level. In university, teachers are free to choose their teaching materials, allowing students to have more opportunities to learn authentic English. Since authentic English is more challenging, students in their first and second years of university are ideally in Stage 3 and work on DDL using authentic English with support from their instructors. Then, in the third and fourth years, students will move on to Stage 4, where they perform DDL autonomously, using authentic English. In Japan, university students aim for Society for Testing English Proficiency (STEP) pre-1st and 1st grade, which correspond to CEFR B1 and B2. Therefore, we can consider Stage 3 as corresponding to the CEFR B1 level, and Stage 4, B2 and above. Although English education begins in the 3rd grade in Japan, students learn English primarily through oral English until 6th grade, with little explicit instruction on English grammar. Therefore, our eDDL tool for elementary school students, which is the sister version of hDDL, would be incorporated into classes at the Pre A1 level at Stage 1.

The above target settings may seem low by world standards. In the Japanese language use environment, exposure to English is limited to English classes at school, and almost all daily life is conducted in Japanese. In other words, students learn English as a foreign language (EFL), so the goal setting is lower than in ESL (English as a Second Language) countries, where English is used inside and outside of school. Furthermore, according to Allan (2009), among authentic materials, English for Specific Purposes (ESP) corpora are easier than general-purpose corpora such as the British National Corpus (BNC) because ESP is more accessible and relevant to students. Learners at the B1 and B2 levels are unlikely to be able to deal with the peripheral linguistic content of general-purpose corpora. Therefore, ESP corpora and general-purpose corpora need to be considered separately among authentic data.

To date, most students we have worked with at secondary schools are at Stage 1 in Figure 1. However, when one of the authors observed an 8th grade DDL class, it was found that students who had become familiar with using DDL tools voluntarily opened the hDDL website to check the superlative when they had questions about English. This suggests that these students are making progress from Stage 1 to Stage 2 on their own.

3.3 Release and Revision of hDDL

The DDL Project began in 2013 and for the first five years, we mostly used paper-based DDL. At that time, few schools had the technological infrastructure to support using DDL, so paper-based DDL was a practical choice. We first released eDDL in April 2019. We subsequently released hDDL in August 2019. Both hDDL and eDDL are available without cost or registration, so they can be accessed anytime by learners or teachers, and both were based on the pedagogical framework of SCoRE but with a unique corpus that matched the English level of the target learners. Additionally, we developed a corpus search tool to search this corpus. In this section, we describe the development of hDDL.

3.3.1 Development of Pedagogical Corpus

hDDL is equipped with its own pedagogical corpus. To create it, we developed a reference corpus that was collected from copyright-free language data, for example, a project called Tatoeba (https://tatoeba.org/eng/). We combined it with search software and named it BES (Basic English Sentence) Search. This software allows users to set search criteria such as sentence length, grammatical item, lemma, part of speech, and/or phrase to search for English sentences suitable for learning objectives and the learners' level. Next, using BES Search as a reference corpus, a team of educators including an author of Japanese government-authorized English textbooks, Japanese English teachers, and native English teachers collaborated to create English sentences, one by one, and their Japanese translations, at the level of secondary school students. The hDDL corpus is a collection of these original sentences.

BES Search (https://bessearch.ddl-study.org/) was later released to the public as an introductory English sentence search software. It has approximately 1,337,000 copyright-free sentences (10,750,000 words). It uses English and Japanese as the language for instruction and for the manual. BES Search is a tool that helps teachers and material developers to create their teaching materials efficiently.

The characteristics of the hDDL pedagogical corpus created in this way are as follows.

- The corpus consists of complete English sentences (rather than partial concordance lines).
- All the sentences are copyright free. Teachers and students can download and use them freely.
- Sentence length, vocabulary, and grammar are level-appropriate.
- Instead of random topics as often found in newspapers and general corpora, we set up similarly-aged fictional characters with biographies including family members, pets, friends, teachers, subjects they are good at and bad at, hobbies, personalities, and so on to create a narrative in the English hDDL sentences. The characters were introduced with illustrations to promote familiarity with the sentences and to make them interesting.
- It is an English–Japanese parallel corpus.
- Pronunciations of all English sentences can be checked.

3.3.2 Development and Revision of Search Subtools

Since its release in August 2019, the hDDL search functions have been revised yearly. These are shown in Table 2. The hDDL corpus has also been expanded every year by increasing the size of the database.

Table 2. History of hDDL Development

Version	Release	Points of modification
1.00	August, 2019	Released the pattern browser search (search by selecting a grammar category)
1.10	September, 2020	Added a concordance search (search by entering a search expression)
1.20	August, 2021	Added an English quiz (sorting questions) Added English as the display language
1.30	October, 2022	Added Auto Search Added a DIY Search

The revision cycle is shown in Figure 2. First, the DDL tool was released, then implemented in schools. Once released, we asked cooperating schools to use the system and then asked students and teachers to comment on any difficulties they had in using hDDL or to suggest any functions they want to see in the next revision. In addition, one of the authors observed students using hDDL in classes. Based on the data collected, further revisions were made.

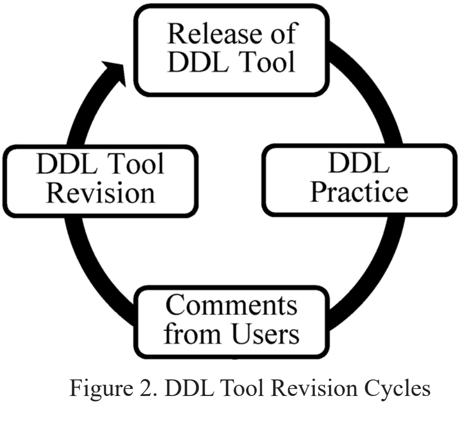
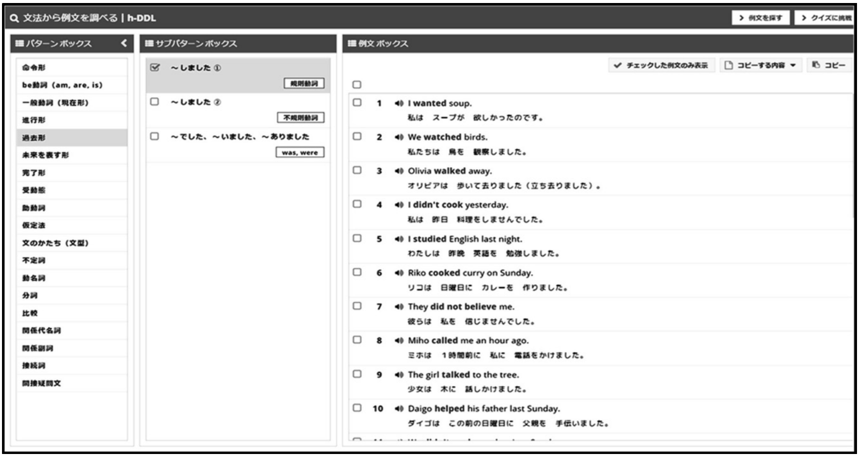


Figure 2. DDL Tool Revision Cycles

In the original release, hDDL included a pattern browser (Figure 3). This allowed students to click on the grammar item they wanted to learn. In Figure 3, the top left column shows the list of grammar items. The second left column shows subcategories of the chosen grammar item. The right column shows the extracted sentences. In this case, students chose "past tense" on the left and chose "regular verbs" in the next column among the choices of "regular verbs," "irregular verbs," and "was, were." In the right column, students can see English sentences and Japanese translations.



□ 1 I **wanted** soup.
私は　スープが　欲しかったのです。

□ 2 We **watched** birds.
私たちは　鳥を　観察しました。

□ 3 Olivia **walked** away.
オリビアは　歩いて去りました（立ち去りました）。

□ 4 I **didn't cook** yesterday.
私は　昨日　料理をしませんでした。

□ 5 I **studied** English last night.
わたしは　昨晩　英語を　勉強しました。

□ 6 Riko **cooked** curry on Sunday.
リコは　日曜日に　カレーを　作りました。

□ 7 They **did not believe** me.
彼らは　私を　信じませんでした。

□ 8 Miho **called** me an hour ago.
ミホは　１時間前に　私に　電話をかけました。

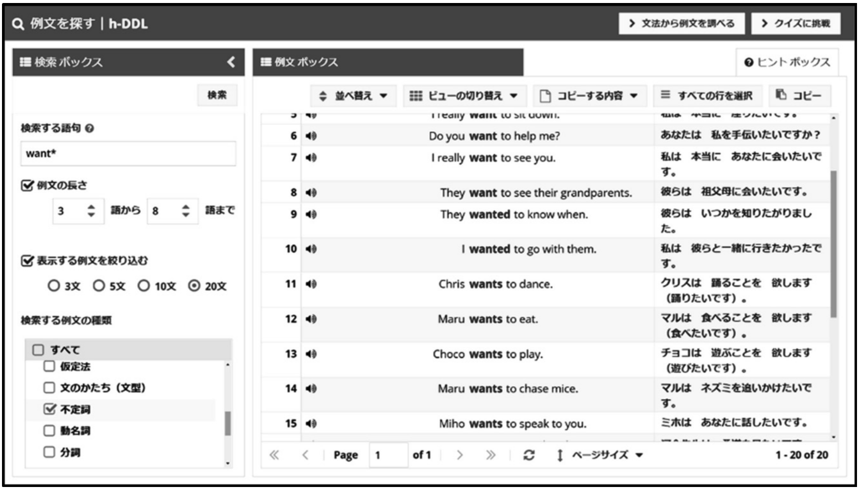
□ 9 The girl **talked** to the tree.
少女は　木に　話しかけました。

□ 10 Daigo **helped** his father last Sunday.
ダイゴは　この前の日曜日に　父親を　手伝いました。

Figure 3. Pattern Browser Search

Using this pattern browser was an easy way to do a search, but students could not search for grammatical items that were not on the list. In order to allow students to search freely and to experience the real pleasure of searching a corpus, we added a new concordance search function in version 1.10. In the top left column in Figure 4, students type in the word or phrase they want to search (e.g., *want*). From the next row, they choose the sentence length (e.g., between 3 and 8 words in the displayed sentences). They can also choose the number of sentences displayed on the screen from 3 to 20 or more. From the bottom left column, they can choose the grammar item to search. In this case, students chose infinitive sentences following *want*, *wants*, and *wanted* ([want*]). With this added function, students can look up English sentence examples for language issues they are curious about but that do not appear in the pattern browser list. However, this subtool requires students to input the search formula by themselves to extract the grammar items they wanted to explore.

Next, in version 1.20, we added English as an instructional language so that hDDL can be used by non-Japanese learners, and created an English quiz for students



検索する語句：want*

☑ 例文の長さ　3 語から 8 語まで

☑ 表示する例文を絞り込む　○ 3文　○ 5文　○ 10文　● 20文

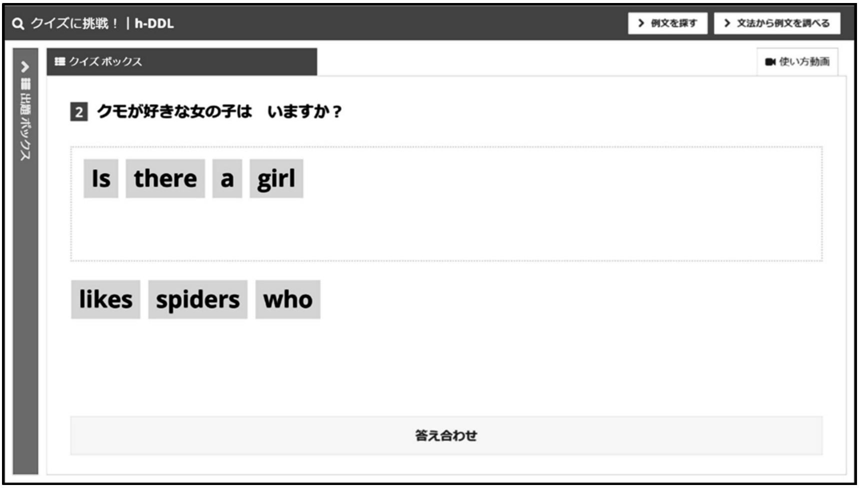
検索する例文の種類
□ すべて
□ 仮定法
□ 文のかたち（文型）
☑ 不定詞
□ 動名詞
□ 分詞

5	I really **want** to sit down.	私は　本当に　座りたいです。
6	Do you **want** to help me?	あなたは　私を手伝いたいですか？
7	I really **want** to see you.	私は　本当に　あなたに会いたいです。
8	They **want** to see their grandparents.	彼らは　祖父母に会いたいです。
9	They **wanted** to know when.	彼らは　いつかを知りたがりました。
10	I **wanted** to go with them.	私は　彼らと一緒に行きたかったです。
11	Chris **wants** to dance.	クリスは　踊ることを　欲します（踊りたいです）。
12	Maru **wants** to eat.	マルは　食べることを　欲します（食べたいです）。
13	Choco **wants** to play.	チョコは　遊ぶことを　欲します（遊びたいです）。
14	Maru **wants** to chase mice.	マルは　ネズミを追いかけたいです。
15	Miho **wants** to speak to you.	ミホは　あなたに話したいです。

Page 1 of 1　　1 - 20 of 20

Figure 4. Concordance Search

to use to check their knowledge of grammar. This is done in the form of "rearrangement questions" in which students look at Japanese sentence construction and arrange the English words to match the Japanese. In the example shown in Figure 5, the student



2 クモが好きな女の子は　いますか？

Is there a girl

likes spiders who

Figure 5. English Quiz

has correctly ordered the first half of the sentence but has not yet done the second set of words. The quiz tool is easy to use and shows scores, so students can answer the questions as if they were doing a game. Best of all, because it uses sorting questions, students will pay attention to the structure of the English sentences.

3.3.3 Troubleshooting

The concordance search subtool added in 2020 (v. 1.10) to support the long-term goal of developing of autonomous corpus users created a problem—students had to type complex search formulas into their computer terminals. When we conducted a questionnaire on the use of hDDL among about 300 7th to 9th grade students and four English teachers who had focused on auxiliary verbs (in the 7th grade), gerunds (in the 8th grade), and present perfect (in the 9th grade), we received the following comments on the concordance search subtool (English translations of the original Japanese):

- I have a hard time using the DDL tool. I needed instructions from the teacher.
- I want the DDL tool to be a little easier to look up.
- I would like the DDL tool to be a little easier to search.

In addition, when one of the authors observed a class, she found that students had difficulty distinguishing between full-width and half-width characters, upper and lower case letters, the areas that require spaces and those that do not, and in using the keyboard.

4. Revisions to Foster Self-Directed Learners

To address this problem of the complicated input of search formulas, we made two revisions. First, “Auto Search” was added to display the concordance line with a single click. Although this subtool was convenient and much easier to use, it did not allow students to see or understand a search formula. Because this is a skill needed for independent corpus users, we added “DIY Search,” which allows students to create and/or customize a search formula—students can see and touch the search formula. These features are described in more detail in the next sections.

4.1 Auto Search

Each hDDL sentence has the information of a grammar item. For example, the sentence *Life has changed* contains information on the identity (ID) of the grammatical item that indicates what it is, in this case [Perfect Forms–Perfective]. Auto Search combines the grammar item with a predefined search formula and displays sentences of a specific grammar item in the Key Word in Context (KWIC) concordance line. In other words, for [Perfect Forms–Perfective], the concordance lines are generated sentences that include this grammar item, with the past participle excluding “been” (expressed as [pos="VBN" & word!="been"] in the search formula of the corpus query language [hereafter CQL]) specified as the keyword. Auto Search is done in two steps with one click. Figure 6 shows an example. First, a student selects the grammar item to search from the top row “❶ Select a grammar item.” Then, in the lower section “❷ Choose a Search Pattern,” the subordinate grammatical items selected in ❶ will be displayed. The grammar item or word the student wants to display is selected from here (Figure 6), and the KWIC concordance (Figure 7) will be displayed in the box to the right. The three dots give the option of going to the DIY Search.

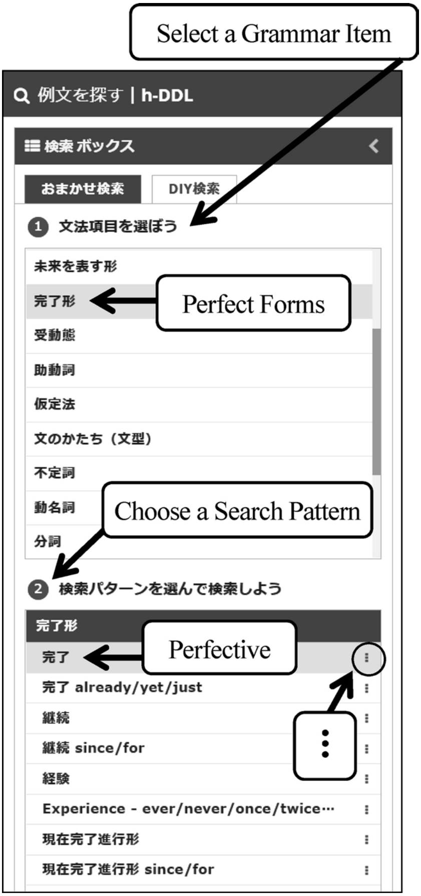
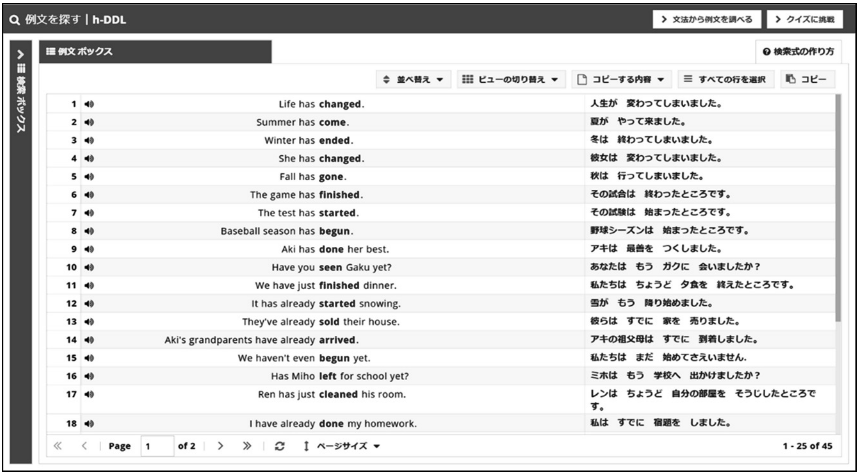


Figure 6. Auto Search

When students previously entered search formulas, they sometimes could not gain search hits correctly because they included extra spaces or unintentionally used full-width alphabets. With Auto Search, only two clicks are needed to display the concordance lines (Figure 7). Thus, students can immediately work on observing and analyzing the example sentences. This reduces the amount of time spent on DDL during class, allowing for more effective use of class time. Thus, with Auto Search, students can use the concordance search to analyze sentences even before they learn the rules of the search formula. They will be able to experience the pleasure of DDL using a corpus.



1	Life has **changed**.	人生が　変わってしまいました。
2	Summer has **come**.	夏が　やって来ました。
3	Winter has **ended**.	冬は　終わってしまいました。
4	She has **changed**.	彼女は　変わってしまいました。
5	Fall has **gone**.	秋は　行ってしまいました。
6	The game has **finished**.	その試合は　終わったところです。
7	The test has **started**.	その試験は　始まったところです。
8	Baseball season has **begun**.	野球シーズンは　始まったところです。
9	Aki has **done** her best.	アキは　最善を　つくしました。
10	Have you **seen** Gaku yet?	あなたは　もう　ガクに　会いましたか？
11	We have just **finished** dinner.	私たちは　ちょうど　夕食を　終えたところです。
12	It has already **started** snowing.	雪が　もう　降り始めました。
13	They've already **sold** their house.	彼らは　すでに　家を　売りました。
14	Aki's grandparents have already **arrived**.	アキの祖父母は　すでに　到着しました。
15	We haven't even **begun** yet.	私たちは　まだ　始めてさえいません.
16	Has Miho **left** for school yet?	ミホは　もう　学校へ　出かけましたか？
17	Ren has just **cleaned** his room.	レンは　ちょうど　自分の部屋を　そうじしたところです。
18	I have already **done** my homework.	私は　すでに　宿題を　しました。

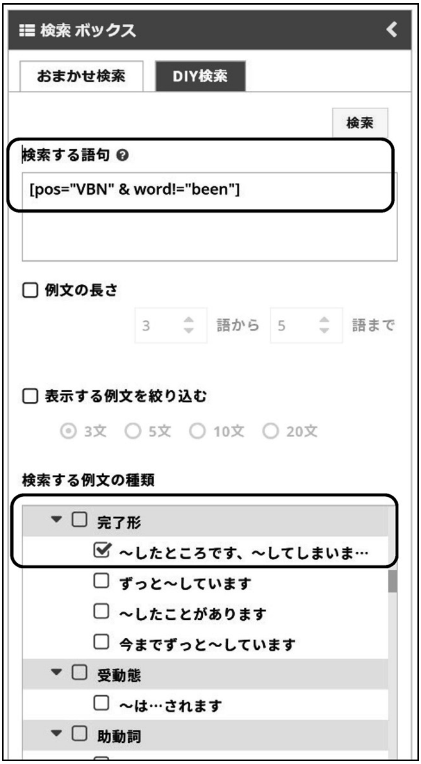
1 - 25 of 45

Figure 7. An Example of the Result of a [Perfect Forms–Perfective] Concordance Search

4.2 DIY Search

Auto Search is a useful feature, but it does not allow students to see, create, or become familiar with the search formula. Therefore, in this modification, we added a DIY search function along with Auto Search. An important feature of the DIY search is the ability to paste the search condition criteria from Auto Search (grammar items and search formulas) into the DIY search screen (see the section marked with the rectangle at the top of Figure 8).

For example, if a student wants to search for sentences in the [Perfect Forms–Perfective], s/he would click the three dots [⋮] to the right of [Perfective] in Figure 6 (see the section with circle) and then click [Set as DIY Search] from its pulldown menu. The screen will then switch to the DIY Search. The search formula selected will be automatically pasted into the "Search for" window. (See the rectangle at the top of Figure 8.) At the same time, a checkmark will automatically be placed in the [Type of



検索 ボックス

おまかせ検索 DIY検索

検索

検索する語句

[pos="VBN" & word!="been"]

□ 例文の長さ

3 語から 5 語まで

□ 表示する例文を絞り込む

● 3文 ○ 5文 ○ 10文 ○ 20文

検索する例文の種類

▼ □ 完了形

☑ ～したところです、～してしまいま…

□ ずっと～しています

□ ～したことがあります

□ 今までずっと～しています

▼ □ 受動態

□ ～は…されます

▼ □ 助動詞

Figure 8. DIY Search

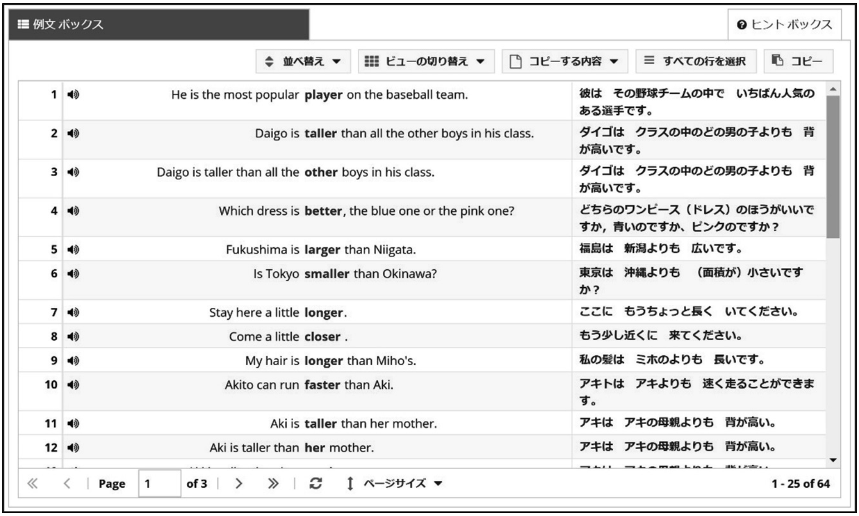
Sentences to Search] box for the meaning of "Perfect Forms–Perfective." (This is indicated by the second rectangular box.) From here, when the student clicks [Search], the same results as in the previous Auto Search will then be displayed.

Students can modify this search formula and change the type and range of example sentences to be searched. In this way, the search range can be widened or narrowed. By modifying an existing search formula, it is easier to understand the rules of the search formula than by creating a search formula from scratch. In addition, there is a tutorial called "Rules for Search Expressions." By using this tutorial, students can learn the rules for CQL search expressions in a single step. In this way, the DIY Search allows students to hone their own corpus search skills.

5. Revisions and Improvements

As a result of these modifications, the hDDL concordance function has been improved in terms of both operability and search accuracy. In the previous concordance function, only surface forms could be specified as search terms for Auto Search. For example, when the conventional concordance function was used to search for comparative classes (*taller*, *other*, *better*, *closer*, etc.), erroneous KWIC words such as *player* and *her* were included in the search, as shown in Figure 9 (*He is the most popular **player** on the baseball team*; *Aki is taller than **her** mother*). The new concordance feature allows the original CQL search to be performed, making the search much more accurate, and this type of noise contamination is minimized. Additionally, the previous concordance feature was limited to searching only comparative classes. However, with the new function, it is now possible to search for more than one adjective before a noun.

Since the release of version 1.30, we have not received any negative feedback about Auto Search from students who used hDDL for the first time. We have, however, received comments from secondary school students and teachers who have used hDDL for some time indicating that it became much easier to use.



1	He is the most popular **player** on the baseball team.	彼は　その野球チームの中で　いちばん人気のある選手です。
2	Daigo is **taller** than all the other boys in his class.	ダイゴは　クラスの中のどの男の子よりも　背が高いです。
3	Daigo is taller than all the **other** boys in his class.	ダイゴは　クラスの中のどの男の子よりも　背が高いです。
4	Which dress is **better**, the blue one or the pink one?	どちらのワンピース（ドレス）のほうがいいですか，青いのですか、ピンクのですか？
5	Fukushima is **larger** than Niigata.	福島は　新潟よりも　広いです。
6	Is Tokyo **smaller** than Okinawa?	東京は　沖縄よりも　（面積が）小さいですか？
7	Stay here a little **longer**.	ここに　もうちょっと長く　いてください。
8	Come a little **closer** .	もう少し近くに　来てください。
9	My hair is **longer** than Miho's.	私の髪は　ミホのよりも　長いです。
10	Akito can run **faster** than Aki.	アキトは　アキよりも　速く走ることができます。
11	Aki is **taller** than her mother.	アキは　アキの母親よりも　背が高い。
12	Aki is taller than **her** mother.	アキは　アキの母親よりも　背が高い。

Figure 9. Noise in the Search Results in the Previous Version

6. Conclusion

Since its inception, the DDL Project has inched forward in bringing DDL into elementary and secondary level EFL classrooms. Over the course of nearly a decade, we have created a simplified pedagogical corpus for the elementary level and another for the secondary school level along with corresponding search tools (eDDL and hDDL). Although COVID-19 created many challenges, one benefit has been that efforts to provide students with “one computer terminal per student” has meant many more computers are now available in schools in Japan. Computer terminals have become as common and convenient for students as pencils and notebooks. With this change in the classrooms, eDDL and hDDL will become more accessible and may contribute to the goal of nurturing future autonomous corpus users. In order to achieve this goal, it is necessary for educators to develop and verify effective DDL instruction for this long-term plan.

Acknowledgment

This research was supported by JSPS KAKENHI Grant Number JP20H01277, Grant-in-Aid for Scientific Research (B). The authors cordially appreciate the insightful comments offered by reviewers in publishing this manuscript, which is a revised version of a presentation at the 48th JAECS research conference 2022.

References

Allan, R. (2009). Can a graded reader corpus provide ‘authentic’ input? *ELT Journal*, *63*(1), 23–32.

Beresova, J. (2015). Authentic materials–Enhancing language acquisition and cultural awareness. *Procedia-Social and Behavioral Sciences*, *192*, 195–204

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

Chujo, K., Oghigian, K., & Akasegawa, S. (2015). A corpus and grammatical browsing system for remedial ESL learners. In A. Lenko-Szymanska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 109–128). John Benjamins.

Crosthwaite, P. (Ed.). (2020). *Data-driven learning for the next generation*. Routledge.

Elmansi, H., Dadour, E., Qoura, A., & Hamada, T. (2021). The impact of data-driven Learning based program on developing student teachers, lexico-grammatical performance skills in EFL writing. *Journal of Research in Curriculum, Instruction and Educational Technology*, *7*(3), 37–65.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ*, *8*, 1–35.

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.

Kakiba, A., Nishigaki, C., & Oghigian, K. (2021). DDL application to the seventh grade EFL classroom in Japan. *Bulletin of the Faculty of Education, Chiba University*, *69*, 179–167.

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, *40*(5), 721–753.

Lin, M. H. (2021). Effects of data-driven learning on college students of different grammar proficiencies: A preliminary empirical assessment in EFL classes. *Sage Open*, *11*(3), 1–15.

Loewen, S. (2020). *Introduction to instructed second language acquisition (Second Edition).* Routledge.

Ministry of Education, Culture, Sports, Science and Technology. (2017). *The Course of study for junior high School* (in Japanese). Higashiyama Shobo.

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

Mull, J., & Conrad, S. (2013). Student use of concordance for grammar error correction. *ORTESOL Journal*, *30*, 5–14.

Nishigaki, C., Akasegawa, S., Kawana, T., Nakai. K., Kenmoku, S., & Yamazaki, T. (2021). Classroom application of a web-based DDL support tool in a secondary school (ISSN 2436-6447). *Proceedings of the JAECS conference 2021*, 97–102. https://jaecs.com/conf_47/jaecs47_proceedings.pdf (Retrieved October 30, 2022)

Nishigaki, C., Akasegawa, S., & Oghigian, K. (2022). Development of an online DDL tool for secondary school learners. *Bulletin of the Faculty of Education, Chiba University*, *70*, 289–297.

Nishigaki, C., Chujo, K., Kamiya, N., Oyama, Y., Abe, N., Monoi, N., & Yokota, A., (2018), Promoting noticing of grammar rules through data-driven learning (in Japanese). *Language Learning and Educational Linguistics 2018–19*, 59–66.

Nishigaki, C., & Kakiba, A. (2023). What does data-driven learning (DDL) bring out in grammar learning? *Bulletin of the Faculty of Education, Chiba University*, *71*. 197–207.

Nishigaki, C., Hoshino, Y., Abe, T., Kamiya, N., Oyama, Y., & Ishii Y. (2020). Development of a data-driven English learning website for elementary school students (in Japanese). *JES Journal*, *20*, 367–382.

Nishigaki, C., Oyama, Y., Kamiya, N., Yokota, A., & Nishizaka, T. (2015). Data-Driven Learning and Focus on Form (in Japanese). *KATE Journal*, *29*, 113–126.

Perez-Paredes, P. (2020). The pedagogic advantage of teenage corpora for secondary school learners. In P. Crosthwaite (Ed.). *Data-driven learning for the next generation* (pp. 67–87). Routledge.

Saeedakhtar, A., Bagerin, M., & Abdi, R. (2020). The effect of hands-on and hands-off data-driven learning on low-intermediate learners' verb-preposition collocations. *System*, *91*, 1–14.

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

Suzuki, H., (2021, July 5). Introduction to deep learning (3): Why "independent learning" is necessary (in Japanese). *Education News*, https://www.kyobun.co.jp/kyosai/k20210705_02/ (Retrieved October 11, 2022)

Tsai, K. J. (2019). Corpora and dictionaries as learning aids: Inductive versus deductive approaches to constructing vocabulary knowledge. *Computer Assisted Language Learning*, *32*(8), 805–826.

Timmis, I. (2015). *Corpus linguistics for ELT: Research and practice.* Routledge.

Vyatkina, N. (2013). Discovery learning and teaching with electronic corpora in an advanced German grammar course. *Die Unterrichtspraxis/Teaching German*, *46*(1), 44–61.

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.

（西垣知佳子　千葉大学）
（赤瀬川史朗　Lago NLP）

「研究ノート」

ESP for Art Majors: Development of the Ballet English Corpus (Ver. 1.0)

Hiroko USAMI

Abstract

English for specific purposes (ESP) examines English used in specific fields, such as business, law, science and technology, medicine, nursing, tourism, and aviation (e.g., Parkinson, 2013; Bargiela-Chiappini & Zhang, 2013; Northcott, 2013; Moder, 2013; Ferguson, 2013). In corpus linguistics, specialised corpora describing specific English have been constructed and applied to teaching (e.g., the Air Traffic Control Complete Corpus by Godfrey (1994) and the Nursing Medical Chart Corpus by Ohashi et al. (2020)). However, English used in the arts, such as music, performing arts, fine arts, and dance, especially classical ballet, has been inadequately examined. Among various kinds of dance, classical ballet has been enjoyed for different purposes by people of all ages in Japan. In addition, an increasing number of Japanese classical ballet dancers have been performing with overseas ballet companies, and the number of adults learning and watching classical ballet has been increasing (e.g., Oyama & Umino, 2022; Showa Academia Musicae Ballet Research, 2010; 2022). This study aims to examine a specialised corpus, the Ballet English Corpus (BEC) (Ver. 1.0), which contains approximately 1 million words from written texts in ten different categories related to classical ballet, which involves people of all ages for different purposes. This study describes the corpus design and analyses the wordlist and keyword list in the BEC. The analysis of the BEC indicates that the word *ballet* is frequently used in the contexts of ballet companies, schools, or dancers' ranks and that some multiword units used typically in ballet repertoires or dancers' biographies are found in n-grams. The keyword list indicates that the word *ballet*, the lemma DANCE, French words, repertoire names, role names, and dancers' ranks occur characteristically and that verb keywords are characteristically used in ballet dancers' biographies. In addition, some keywords with specific meanings, such as *principal*, *role*, and *company*, can be found only in classical

ballet contexts.

1. Introduction

English for specific purposes (ESP) can be defined as ‘an approach to language teaching that targets the current and/or future academic or occupational needs of learners, focuses on the necessary language, genres, and skills to address these needs, and assists learners in meeting these needs through the use of general and/or discipline-specific teaching materials and methods’ (Anthony, 2018:1). The ESP is concerned with English used in various specific fields, such as business, law, science and technology, medicine, nursing, tourism, and aviation (e.g., Parkinson, 2013; Bargiela-Chiappini & Zhang, 2013; Northcott, 2013; Moder, 2013; Ferguson, 2013). An increasing number of ESP textbooks have been published and made available for English teachers to use when teaching specific English to university students or advanced learners with different majors. Among various kinds of corpora in the field of corpus linguistics, specialised corpora, describing specific fields of English, have been constructed and can be applied to analysing frequently used vocabulary and grammar; compiling words, phrases, and grammar lists; revising textbooks; writing question items testing technical terms; and developing teaching platforms (e.g., the Air Traffic Control Complete Corpus by Godfrey (1994) and the Nursing Medical Chart Corpus by Ohashi et al. (2020)). In the fine arts, a specialised corpus containing three different sub-lists (painting, sculpture, and graphic arts) has been constructed (Chatburapanun & Yordchim, 2014). Specialised corpora for dance, specifically for dance notation, have also been constructed (e.g., Bull, 1996; Essid et al., 2012). However, corpora describing the use of English in classical ballet situations remain limited.

Classical ballet has been enjoyed for different purposes by people of all ages in Japan. Young Japanese classical ballet dancers have enrolled in overseas classical ballet schools in increasing numbers to be trained as professional classical ballet dancers. In addition, overseas ballet companies have hired more Japanese classical ballet dancers to perform professionally (e.g., the Association of Japanese Ballet Companies, 2018; 2019; 2020; 2021; 2022). Classical ballet is also popular among non-professional classical ballet learners in Japan, who enjoy it by, for example, learning it as a hobby, doing it as a form of exercise, or simply watching performances (e.g., Oyama & Umino, 2022; Showa Academia Musicae Ballet Research, 2010; 2022).

This study aims to describe a specialised corpus created by the author, the Ballet English Corpus (BEC), which can be applied to the study of the specific field of English used in classical ballet contexts and could be helpful for both professional and non-professional classical ballet learners and dancers in Japan. The BEC (Ver. 1.0) currently contains approximately 1 million words from written texts used in ten different contexts related to classical ballet (ballet techniques, companies, studios, history, schools, theatres, people, narratives, repertoires, and miscellaneous), which involves people of various ages for different purposes. This study provides details of the corpus design and analysis of the n-grams and keyword list in the BEC.

2. Literature Review

2.1 English Textbooks and Specialised Corpora for English for Art

There are many ESP textbooks in circulation targeting various genres. In fact, according to this survey examining ESP textbooks available in Japan in 2022 and published by 11 Japanese and overseas publishers for university English, 11 ESP textbooks for art, 12 ESP textbooks for music, and as many as 45 ESP textbooks for movies have been published. However, most ESP textbooks for music teach English in restricted contexts, such as understanding song lyrics.

In addition to ESP textbooks strictly available for university teachers and students, several English conversation textbooks have been published. They can be accessed by anyone involved with or interested in music, such as music majors in high school and university (Kubota & Orui, 2014, 2017; Kanda et al., 2016). These textbooks help those interested in learning English conversation in music learning contexts, such as music lessons, accompanying musical instruments, and concerts. However, only one English conversation textbook has recently been published for dance and theatre studies, especially classical ballet, targeting only children interested in learning classical ballet overseas. The textbook aims to help them learn English in the form of conversations by providing valuable words and phrases that they can use in various contexts related to classical ballet, such as when attending classes, watching performances, shopping for dancewear in ballet stores, and participating in school activities (Ito, 2021).

In the corpus linguistic field, specialised corpora describing specific fields of English have been constructed. Examples of such corpora include the Air Traffic Control Complete Corpus by Godfrey (1994) and the Nursing Medical Chart Corpus by Ohashi et al. (2020). Nesi (2013) points out that ESP corpora can be utilised in teaching, such as examining and analysing concordance lines and obtaining wordlists, keywords, and n-grams for teaching materials as well as test design and data-driven learning (e.g., Yang, 1986; Harwood, 2005). Regarding the fine arts, a specialised corpus containing three different sub lists (painting, sculpture, and graphic arts) has been constructed and applied to ESP, especially in creating teaching and learning platforms (Chatburapanun & Yordchim, 2014). Specialised corpora, specifically for dance notation, have also been constructed (e.g., Bull, 1996; Essid et al., 2012).

2.2 Popularity of Classical Ballet in Japan

Within the performing arts field and among various kinds of dance, classical ballet has been especially popular in Japan across different age groups. According to a survey conducted every five years (in 2011, 2016, and 2021) by Showa Academia Musicae Ballet Research (2022), approximately 256,000 people, 0.20% of the total Japanese population, learned classical ballet in Japan in 2021. Unlike other countries, Japan has no national classical ballet schools or universities. Therefore, most classical ballet learners learn at studios run privately (71.6%) or by companies (22.3%) (Oyama & Umino, 2022; Showa Academia Musicae Ballet Research, 2022). Showa Academia Musicae Ballet Research (2022) surveyed the ages of students at classical ballet studios. More than 80% of classical ballet studios teach children in elementary school or younger, while the number of classical ballet studios that teach junior and high school students is slightly lower. More than 70% of classical ballet studios teach adults in their 40s and 50s. Comparing the population of children and adults, that of classical ballet learners under three years of age and over 70 years of age is much lower than the rest of the age groups. Therefore, classical ballet can be learned by people of all ages, from children aged four to senior citizens under 70.

Some classical ballet learners have been practising classical ballet since they were children to become professional dancers. More and more children typically learn classical ballet in Japan until junior high school, and then transfer to an overseas classical ballet school. They adopt this strategy because no nationally qualified classical ballet schools or universities exist in Japan and they would like to be hired and perform with overseas ballet companies after graduating. Approximately 13% of the classical ballet studios have graduates who dance with overseas classical ballet schools and companies. They start their career as professional dancers in their late teens or early twenties, either at overseas or Japan-based ballet companies. (Oyama & Umino, 2022; Showa Academia Musicae Ballet Research, 2022). Therefore, young classical ballet learners must prepare their resumes and application forms and then audition in English to apply to overseas ballet schools or learn classical ballet (e.g., simply taking classical ballet lessons or studying its background) overseas at English-speaking classical ballet schools.

The number of Japanese classical ballet dancers working in overseas classical ballet companies has gradually increased between the 2017–2018 and 2021–2022 seasons (the Association of Japanese Ballet Companies, 2022). Therefore, adult professional classical ballet dancers must also design their resumes and biographies and audition at overseas classical ballet companies in English. Once they get hired at these classical ballet companies, they take classical ballet lessons and attend rehearsals in English. Sometimes they discuss the repertoires or roles they have performed or will perform with their colleagues in English.

Regardless of age, most classical ballet learners and dancers enjoy classical ballet by watching performances. Non-professional classical ballet learners take classical ballet lessons as a hobby or for exercise (Showa Academia Musicae Ballet Research, 2010). Furthermore, most non-professional classical ballet learners might want to watch classical ballet performances performed by overseas ballet companies or take classical ballet classes or private lessons from international classical ballet teachers in Japan or in foreign countries.

3. Ballet English Corpus (Ver. 1.0)

Increasing numbers of young classical ballet learners and dancers are enrolling in overseas classical ballet schools or performing with overseas classical ballet companies and using English for auditions, class lessons, and rehearsals. Furthermore, non-professional classical ballet learners and dancers are taking lessons in English or watching classical ballet performances in foreign countries. Therefore, the author constructed a specialised corpus—the BEC—so that English texts collected from around the world and generally used in the classical ballet contexts can be examined.

The first step involved developing the Ver. 1.0 of the BEC started with constructing ten different categories with which professional and non-professional classical ballet learners and dancers of all ages would likely be familiar (shown in Table 1). These categories were based on the book *Ballet: The Definitive Illustrated Story* (Durante, 2018), which contains a wealth of information as well as books and websites on classical ballet. As many books and websites as possible written in English on classical ballet, which are accessible only in Japan were collected based on the categories. All the words in the books excluding tables of contents, indexes, glossaries, pictures, and photos were manually typed into text files. Subsequently, the pages and chapters of the same books and websites were categorised into appropriate text files according to the categories.

Table 1. Total numbers of words across ten categories in the BEC (Ver. 1.0)

Category	Ballet basics	Company	Dance studio	History	Narrative	People	Repertoire	School	Theatre	Miscellaneous	Total
Total	54,336	179,300	18,552	4,888	13,432	427,056	228,863	44,965	7,635	39,051	1,018,078

In Table 1, the category ‘ballet basics’ contains texts on basic classical ballet techniques taught in bar and centre lessons and other dance forms, such as character dances and *pas de deux* (‘dance for two’). The category ‘repertoire’ contains texts on classical ballet performances such as *Swan Lake*. The categories ‘company’, ‘dance studio’, ‘school’, and ‘theatre’ contain texts on professional ballet companies, dance studios where anyone can enjoy dancing, ballet schools to train classical ballet dancers, and theatres where classical ballet is performed, respectively. The category ‘history’ contains texts on classical ballet history, while the category ‘narrative’ contains stories related to classical ballet for children. The category ‘people’ contains texts on biographies of known figures related to classical ballet, both current and retired (e.g., classical ballet dancers, choreographers, directors, teachers, and ballet music composers). ‘Miscellaneous’ contains texts related to classical ballet that cannot be classified into the above categories (e.g., costume designs, pointe shoes, lightning, and stage).

By far, ‘people’ has the largest number of words obtained primarily from websites, accounting for almost half the total. The second and third largest categories are ‘repertoire’ and ‘company’, respectively. The categories ‘ballet basics’, ‘school’, and ‘miscellaneous’ are small, and ‘history’ and ‘theatre’ are even smaller. While the ratio of each category is biased, the result of the initial step in developing BEC (Ver. 1.0) (collecting the largest possible number of English books and websites accessible only in Japan) would be a relatively realistic representation.

The software package, #LancsBox (http://corpora.lancs.ac.uk/lancsbox/index.php), which was developed at Lancaster University, was used to analyse the BEC (Ver. 1.0). The Words (for wordlists), Ngrams (for n-grams), and KWIC (for concordances) functions available in #LancsBox were used in this research.

Using the BEC (Ver. 1.0), this study examines what words, multiword units, and keywords are generally used and are helpful for classical ballet learners and dancers regardless of their ages and whether they are professional or non-professional. The research questions addressed in the following sections are: Which 1-, 2-, 3-, and 4-grams frequently occur, and what keywords are found in the BEC (Ver. 1.0)?

4. Analysis

4.1 N-grams in the BEC

First, in order to examine which single words and multiword units are frequently used in the BEC (Ver. 1.0), 1-, 2-, 3-, and 4-grams were obtained using #LancsBox. As Table 2 shows, articles (*the* and *a*), a conjunction (*and*), and prepositions (*in*, *of*, *to*, *with*, *for*, *as*, *at*, *on*, and *by*) frequently occur under the list of 1-grams, and multiword units composed of a preposition or conjunction + article (*of the*, *in the*, *at the*, *to the*, *and the*, *with the*, *for the*, *as a*, and *on the*) frequently occur under the list of 2-grams. In addition, *be* verbs (*was* and *is*) and second and third personal pronouns, such as *her*, *she*, *you*, and *he*, with slightly more third personal pronouns of female (*her* and *she*), appear frequently.

As expected, words related to classical ballet occur: *ballet* is ranked 7th and *dance* 18th. The word *ballet* appears in 2-, 3-, and 4-grams as the names or part of the names of ballet companies or schools (*royal ballet*, *ballet school*, *national ballet*, *the royal ballet*, *english national ballet*, *royal ballet school*, *the royal ballet school*, *new york city ballet, of the royal ballet*, *at the royal ballet*, *the national ballet of*, *national* *ballet of canada*, and *school of american ballet*), and dancers' ranks (*corps de ballet*, *the corps de ballet*, *corps de ballet in*, *dancer corps de ballet*, and *ballet dancer corps de*). The French word *de* is used as the names or part of the names of steps (*pas de* and *pas de deux*) and dancers' ranks (*corps de ballet*, *the corps de ballet*, *corps de ballet in*, *dancer corps de ballet*, *of the corps de*, and *ballet dancer corps de*). Multiword units more specific to ballet used as the names or part of the names of the repertoire occur (*the nutcracker*, *swan lake*, *the sleeping beauty*, and *romeo and julie*t). Therefore, professional and non-professional classical ballet learners and dancers would generally be likely to encounter these words and multiword units in classical ballet situations.

Other multiword units used specifically in dancers' biographies can be found, indicating age (e.g., *the age of*, *at the age*, and *at the age of*), membership (e.g., *a member of the*, *member of the corps*, and *as a member of*), and ticket information (e.g., *more tickets & info* and *tickets & info July*) as follows:

> She began dance classes at the age of three. (BEWWAP118)
> He joined the Company as a member of the corps de ballet in November 2017. (BEWWAP382)
> Upcoming Performances July 1, 2021 8:45 pm Indestructible Light MORE TICKETS & INFOR July 4, 2021 8:00 pm. (BEWWAP060)

In addition, the multiword unit *was born in*, followed by the birthplace of dancers or company staff, and the multiword unit *was promoted to* occur followed by the dancers' rank, such as principal, soloist, or artist. One more characteristic multiword unit *under the direction of* occurs followed by teachers, mothers, directors, or choreographers, as seen below:

> She joined Boston Ballet that year and was promoted to soloist in 2001 and principal in 2003. (BEWWAP404)
> She performed as Clara in Brian Reeder's The Nutcracker under the direction of Ethan Stiefel. (BEWWAP635)

As analysed above, their age, dancers' ranks, ticket information, hometown, promotion, teachers, directors, and choreographers tend to be specifically mentioned in

Table 2. 1-grams, 2-grams, 3-grams, and 4-grams in the BEC

Rank	1-gram	Freq.	2-gram	Freq.	3-gram	Freq.	4-gram	Freq.
1	the	56,264	of the	5,625	written website people	1,567	at the age of	457
2	and	33,624	in the	4,564	the royal ballet	1,431	the corps de ballet	368
3	in	26,660	at the	4,123	corps de ballet	1,026	more tickets & info	358
4	of	21,668	the royal	2,514	the sleeping beauty	1,000	the royal ballet school	297
5	to	20,624	to the	2,378	pas de deux	803	corps de ballet in	281
6	a	18,722	royal ballet	2,224	royal opera house	655	a member of the	272
7	ballet	18,305	and the	2,208	the royal opera	597	new york city ballet	267
8	with	8,732	with the	1,915	english national ballet	597	dancer corps de ballet	258
9	for	8,605	written website	1,884	romeo and juliet	532	of the corps de	249
10	as	7,845	for the	1,741	the age of	531	tickets & info july	248
11	at	7,761	as a	1,669	written book child	476	member of the corps	240
12	was	6,633	ballet school	1,572	at the age	470	as a member of	223
13	is	6,032	website people	1,567	new york city	470	of the royal ballet	215
14	her	5,853	the nutcracker	1,411	royal ballet school	467	at the royal ballet	204
15	she	5,748	pas de	1,281	was born in	454	ballet dancer corps de	195
16	you	5,538	ballet in	1,248	at the royal	453	the national ballet of	189
17	he	5,496	ballet dancer	1,237	of the royal	441	under the direction of	181
18	dance	5,044	on the	1,201	as well as	436	the company as a	180
19	on	4,918	national ballet	1,186	was promoted to	430	national ballet of canada	173
20	by	4,666	swan lake	1,167	one of the	426	school of american ballet	172

dancers' biographies. Therefore, these multiword units would be worth presenting and teaching to young classical ballet learners, dancers, and professional classical ballet dancers who must write their own biographies in English.

4.2 Keyword List in the BEC

Next, in order to examine which single words are characteristically used in the BEC, a keyword list was obtained, comparing the word list of the BEC with that of the BNC2014-baby containing 5,024,072 tokens of written and spoken British English from the period 2010–2017 stored in the #LancsBox.

The word *ballet,* by far, is characteristically used with the highest keyness, approximately five times more than the second word in the rank. Compared with Table 2, the lemma DANCE is used variously as different parts of speech and verb forms (*dance*, *dancer*, *dancers*, *danced*, and *dancing*). Similarly, diverse words related to ballet companies or schools (*royal*, *opera*, *company*, *theatre*, and *school*) and more varied French words (*de*, *pas*, *corps*, *la*, and *deux*) occur as keywords. Furthermore, different words used as the name or part of the names of a repertoire occur as keywords; *nutcracker* from *The Nutcracker*, *swan* and *lake* from *Swan Lake*, *giselle* from *Giselle*, *romeo* and *juliet* from *Romeo and Juliet*, *beauty* and *sleeping* from *The Sleeping Beauty,* and *cinderella* from *Cinderella*. Those related to roles or characters classical ballet dancers portray (*prince, roles, fairy*, and *role*) and the rank of classical ballet dancers in the company (*principal*, *soloist*, and *artist*) occur as keywords.

Table 3. Keyword list in the BEC

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

Furthermore, verb keywords used specifically in dancers' biographies, such as *joined*, *performed*, and *featured*, can be found. For example, the word *joined* tends to be followed by a year (sometimes preceded by a colon), a company name, and a dancers' rank, the word *performed* tends to be followed by a date, a role, or *with* + a company name, and the word *featured* tends to be followed by a preposition + media or by the word *role*, constructing the multiword unit *featured role*, as follows:

Born: Toledo, Ohio Trained: Canada's National Ballet School Joined: 2014

(BEWWAP970)
He joined Birmingham Royal Ballet in 2007. (BEWWAP113)
She joined as an apprentice in 2011. (BEWWAP749)
He performed the role of Prince in The Nutcracker for two years. (BEWWAP697)
She performed with the New York City Ballet in George Balanchine's The Nutcracker. (BEWWAP778)
She was featured in Dance Magazine in 2018. (BEWWAP785)
She created a featured role in Dream within a Dream (deferred). (BEWWAP048)

Therefore, the information on when they joined the company, mentioning the dancers' rank, which role they performed when and in which company, and in which media the dancer was featured, using key verbs such as *joined*, *performed*, and *featured*, would be likely included specifically in dancers' biographies.

In addition, some keywords used with meanings specific only to classical ballet contexts would be worth presenting to classical ballet learners and dancers. For example, the word *principal*, generally the head teacher in school, refers to top-ranked dancers in classical ballet English contexts. The word *role* refers to the role or character that classical ballet dancers perform and portray in classical ballet repertoires. The word *company*, meaning an office in usual contexts, indicates a ballet company. Therefore, learning the meanings of words used specifically in classical ballet contexts will be valuable for classical ballet learners and dancers.

5. Conclusion and Future Research

This paper described the corpus design and briefly analysed a specialised corpus, the version 1.0 of the BEC. Professional and non-professional dancers and learners enjoy classical ballet for different purposes. Therefore, the author constructed the BEC to examine English texts collected used in classical ballet situations worldwide. Currently, the BEC (Ver. 1.0) contains approximately 1 million words grouped into ten categories, with most words classified under 'people'.

The analysis of the 1-, 2-, 3-, and 4-grams indicates that articles, a conjunction, prepositions, and *be* verbs, as well as second and third personal pronouns, frequently occur. Words related to classical ballet (*ballet* and *dance*) appear frequently. Multiword units indicating ballet companies, schools, or dancers' ranks with the word *ballet*, and ballet step names or dancers' ranks with the French word *de* occur frequently. Specifically in dancers' biographies, multiword units concerning age, membership, dancers' ranks, ticket information, hometown, promotion, teachers, directors, and choreographers tend to be presented frequently.

The keyword list indicates that the word *ballet* appears with the highest keyness, and the word *dance* is used in different parts of speech and verb forms. Furthermore, more varied words of ballet companies or schools and French words occur as keywords. Words concerning roles or characters portrayed by classical ballet dancers and those concerning their ranks in the company occur as keywords. In addition, specifically in dancers' biographies, verbs *joined*, *performed*, and *featured* occur characteristically as keywords, followed by specific words. Furthermore, some keywords used with specific meanings only in classical ballet contexts can be found.

At the moment, the BEC (Ver. 1.0) currently contains 1,018,078 words only from written texts used in classical ballet situations. In the future, the BEC will be expanded by adding spoken texts transcribed from recorded classical ballet classes and rehearsals and written texts from more books, pamphlets, or leaflets in theatres which can also be made available for purchase by those buying tickets to the show, sometimes referring to the *Ballet Archive* developed by Showa Academia Musicae (https://ballet-archive.tosei-showa-music.ac.jp).

In future research, the BEC (Ver. 1.0) can be used to teach English to art majors. The BEC consists of ten different categories, and each category should have specific features. Words or multiword units frequently used in each category can be identified for different purposes. Based on analyses of words, multiword units, and keyword lists, vocabulary lists and textbooks can be compiled for young classical ballet learners who intend to learn classical ballet overseas, as well as professional classical ballet dancers working at overseas ballet companies.

Acknowledgements

I would like to thank Editage (www.editage.com) for English language editing.

References

Anthony, L. (2018). *Introducing English for specific purposes*. Routledge.

Bargiela-Chiappini, F. & Zhang, Z. (2013). "Business English." In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes*. Wiley Blackwell, pp. 193–212.

Bull, A. (1996). *The formal description of aerobic dance exercise: a corpus-based computational linguistics approach*. Department of Computer Studies, University of Leeds.

Chatburapanun, J. & Yordchim, S. (2014). "A corpus-based study of fine arts English vocabulary." *UMT Poly Journal Vol. 11* No.2: 6–10.

Durante, V. (2018). *Ballet: the definitive illustrated story*. DK.

Essid, S., Lin, X., Gowing, M., Kordelas, G., Aksay, A., Kelly, P., Fillon, T., Zhang, Q., Dielmann, A., Kitanovski, V., Tournemenne, R., Masurelle, A., Izquierdo, E., O'Connor, N. E., Daras, P., & Richard, G. (2012). "A multi-modal dance corpus for research into interaction between humans in virtual environments." *Journal on Multimodal User Interfaces, 7* (1–2): 157–170.

Ferguson, G. (2013). "English for medical purposes." In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes*. Wiley Blackwell, pp. 243–262.

Godfrey, J. (1994). *Air traffic control complete LDC94S14A*. Linguistic Data Consortium. Retrieved from https://catalog.ldc.upenn.edu/LDC94S14A.

Harwood, N. (2005). "What do we want EAP teaching materials for?" *Journal of English for Academic Purposes*, *4*: 149–161.

Ito, Y. (2021). *Ballet eikaiwa* [Ballet English conversation]. Shinsyokan.

Kanda, H., Horikawa, M. & Horikawa, J. (2016). *Suguni yakudatsu popps eikaiwa – gyoukai yougo mo wakaru! Popular, jazz music no phrase book*. [Useful English conversation – understanding pop music jargon! Phrase book for popular and jazz music]. Stylenote.

Kubota, K. & Orui, T. (2014). *Ondaisei ongakuka no tameno eigo de stepup – ongaku ryugaku de yakudatsu eikaiwa 50 scene*. [Step up in English for music major students and musicians – English conversations in 50 scenes that are helpful in music study abroad]. Stylenote.

Kubota, K. & Orui, T. (2017). *Kaitei ban Ondaisei ongakuka no tameno eigo de step up – ongaku ryugaku de yakudastu eikaiwa 50 scene*. [Step up in English for music major students and musicians – English conversations in 50 scenes that are helpful in music study abroad – revised edition]. Stylenote.

Moder, C. L. (2013). "Aviation English." In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes*. Wiley Blackwell, pp. 227–242.

Nesi, H. (2013). "ESP and corpus studies." In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes*. Wiley Blackwell, pp. 407–421.

Northcott, J. (2013). "Legal English." In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes*. Wiley Blackwell, pp. 213–226.

Ohashi, Y., Katagiri, N., Oka, K., & Hanada, M. (2020). "ESP corpus design: compilation of the Veterinary Nursing Medical Chart Corpus and the Beterinary Nursing Wordlist." *Corpora, 15,* 2: 125–140.

Oyama, K. & Umino, B. (2022). *Nihon no ballet kyouiku kannkyo no jittai bunseki. Ballet kyoiku ni kansuru zenkoku chosa 2021 – kihon houkoku*. [Analysis of the ballet education situation in Japan: Nationwide survey on ballet education – basic report]. Showa Academia Musicae.

Parkinson, J. (2013). "English for science and technology." In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes*. Wiley Blackwell, pp.155–173.

Showa Academia Musicae Ballet Research. (2010). *Nihon no ballet kyoiku ni tekishita ballet kyojyuhou kennkyu –kannren jigyou houkokusyo.* [Research on ballet education methods suitable for ballet education in Japan– Report on related works]. Showa Academia Musicae.

Showa Academia Musicae Ballet Research. (2022). *Nihon no ballet kyoiku ni kansuru zenkoku chosa* – hokokusyo. [Nationwide survey on ballet education in Japan – Report]. Showa Academia Musicae.

The Association of Japanese Ballet Companies. (2018). *Sinshin ballet dancer ikusei narabini ballet dan unei no kiso seibi oyobi seisaku jinnzai ikusei hokokusyo*. [Nurturing of emering ballet dancers and basic maintenance of managing ballet companies and production human resources training – report].

The Association of Japanese Ballet Companies. (2019). *Sinshin ballet dancer ikusei narabini ballet dan unei no kiso seibi oyobi management jinnzai ikusei hokokusyo*. [Nurturing of emerging ballet dancers and basic maintenance of managing ballet companies and management human resources training – report].

The Association of Japanese Ballet Companies. (2020). *Sinshin ballet dancer ikusei narabini ballet dan unei no kiso seibi oyobi management jinnzai ikusei hokokusyo*. [Nurturing of emerging ballet dancers and basic maintenance of managing ballet companies and management human resources training – report].

The Association of Japanese Ballet Companies. (2021). *Sinshin ballet dancer ballet dan unei staff no ikusei, narabini jisedai no kankyaku ikusei ni muketa tyousha hokokusyo*. [Nurturing of emerging ballet dancers and ballet company staff, and survey regarding nurturing of next generation audience – report].

The Association of Japanese Ballet Companies. (2022). *Sinshin ballet dancer ballet dan unei staff no ikusei, narabini jisedai no kankyaku ikusei ni muketa tyousha hokokusyo*. [Nurturing of emerging ballet dancers and ballet company staff, and survey regarding nurturing the next generation audience – report].

Yang, H. Z. (1986). "A new technique for identifying scientific/technical terms and describing science texts." *Literary and Linguistic Computing*, *1*: 93–103.

（宇佐美裕子　東海大学）

英語コーパス学会　第48回大会　プログラム

日　　時　　2022年10月1日（土）9:30–17:40　Zoom開催
開 会 式　　9:30–9:40

ワークショップ：9:40–10:40

ワークショップ1【Running a Vocabulary Course With Lextutor】（場所：ルームA）
講　師：Dr. Tom Cobb (Université du Québec à Montréal)
ワークショップ2【言語データを対象としたKH Coderの活用法】（場所:ルームB）
講　師：樋口　耕一（立命館大学）
ワークショップ3【初めてのXML】（場所：ルームC）
講　師：永崎　研宣（一般財団法人人文情報学研究所）

●研究発表第1セッション　言語資源開発（場所：ブレイクアウトルーム1）
司会：島津美和子（立教大学）

研究発表1：10:50–11:10
The pre-processing of YouTube transcripts for corpus-based spoken language analysis
Christopher Cooper（Rikkyo University）

研究発表2：11:10–11:30
日英・英日パラレルコーパス検索ツール『パラレルリンク』（Ver.1.20）
—インターフェース，検索機能，活用研究などについて—
仁科　恭徳（神戸学院大学）
赤瀬川史朗（Lago NLP）

研究発表3：11:30–11:50
PEPコーパスプロジェクト：その設計と射程
神原　一帆（京都外国語大学·立命館大学）
木村　修平（立命館大学）
近藤　雪絵（立命館大学）
山下　美朋（立命館大学）
山中　　司（立命館大学）

●研究発表第2セッション　文法・統語（場所：ブレイクアウトルーム2）
司会：森下　裕三（桃山学院大学）

研究発表 4：10:50-11:10

受動文における様態副詞の生起位置に関する一考察

西村　知修（石川工業高等専門学校）

研究発表 5：11:10-11:30

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

藏薗　和也（神戸学院大学）

研究発表 6：11:30-11:50

学習英文法における as best as possible の位置付け

松田　佑治（関西外国語大学大学院生）

●研究発表第 3 セッション　英語教育・中高生（場所：ブレイクアウトルーム 3）

司会：田畑　圭介（神戸親和女子大学）

研究発表 7：10:50-11:10

A longitudinal study of fluency based on learner corpus of English conversations

Maxim Tikhonenko（Tokyo University of Foreign Studies, Graduate Student）
Keiko Mochizuki（Tokyo University of Foreign Studies）

研究発表 8：11:10-11:30

日本人中学生英語学習者の take の使用と語義指導への提言

澤口　　遼（関西大学第一中学校）

研究発表 9：11:30-11:50

学習者コーパスの品詞連鎖分析による母語の英語構文への影響

山口　一華（東京外国語大学大学院生）
投野由紀夫（東京外国語大学）

●研究発表第 4 セッション　ESP（場所：ブレイクアウトルーム 4）

司会：中谷　安男（法政大学）

研究発表 10：10:50-11:10

研究論文における英語と日本語のアブストラクトは読み手を意識しているか

浅野　元子（大阪医科薬科大学）
松田　紀子（近畿大学）

研究発表11：11:10-11:30
高校生のライティングに対する添削者の添削エラー分析
—時制の誤りに焦点を当てて—
竹下　綾音（九州大学大学生）

研究発表12：11:30-11:50
Development of the Ballet English Corpus (Ver. 1.0) for art major students
Hiroko Usami（Tokai University）

会員総会　11:50-12:20（場所：メインルーム）

〈休憩　12:20-13:10〉

●研究発表第5セッション　分析手法（場所：ブレイクアウトルーム1）
司会：西村　知修（石川工業高等専門学校）

研究発表13：13:10-13:30
COCAにおける正誤表現と非標準表現について
田畑　圭介（神戸親和女子大学）

研究発表14：13:30-13:50
CasualConc 3.0 - Universal Dependencyタグを利用した文法検索の試み
今尾　康裕（大阪大学）

研究発表15：13:50-14:10
量的概念分析再考：動詞explainを例に
菅原　裕輝（大阪大学）
神原　一帆（京都外国語大学・立命館大学）

●研究発表第6セッション　英語学・社会言語学（場所：ブレイクアウトルーム2）
司会：藏薗　和也（神戸学院大学）

研究発表16：13:10-13:30
米国一般教書演説に出現する分離不定詞の効果に関する予備的研究
福本　広光（大阪大学大学院生）

研究発表17：13:30-13:50
男性リーダーの発話データの特徴分析：オックスフォード・ユニオン及びTED Talk分析の示唆
中谷　安男（法政大学）

研究発表18：13:50-14:10

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

島津美和子（立教大学）

●研究発表第7セッション　英語教育・大学生（場所：ブレイクアウトルーム3）

司会：浅野　元子（大阪医科薬科大学）

研究発表19：13:10-13:30

日本人英語学習者の「うなぎ文」の使用に関する分析

藤原　康弘（名城大学）
岩男　考哲（神戸市外国語大学）
伊藤　　創（関西国際大学）
仲　　　潔（岐阜大学）

研究発表20：13:30-13:50

The use of nominalization features in the academic texts by Japanese learners

Kaede Hanawa（Tokyo University of Foreign Studies, Graduate Student)

Yukio Tono（Tokyo University of Foreign Studies)

研究発表21：13:50-14:10

The misuses of English articles in compositions of L1 Chinese and Japanese learners: A corpus-based study

Xiao Sun（Kyoto Sangyo University, Graduate Student)

●研究発表第8セッション　DDL（場所：ブレイクアウトルーム4）

司会：水本　　篤（関西大学）

研究発表22：13:10-13:30

初等・中等教育向け DDL ツールが目指す「自律的なコーパスユーザーの育成」

西垣知佳子（千葉大学）

赤瀬川史朗（Lago NLP）

研究発表23：13:30-13:50

中学校通常英語授業における DDL の活用を目指して：学習環境・タイミングの違いによる比較

中井　康平（千葉大学教育学部附属中学校）

水本　　篤（関西大学）

西垣知佳子（千葉大学）

研究発表24：13:50-14:10
英作文における動詞―名詞コロケーション産出に対するDDLの効果
佐竹　由帆（青山学院大学）

●研究発表第9セッション　分析手法・統計（場所：ブレイクアウトルーム1）
司会：今尾　康裕（大阪大学）

研究発表25：14:20-14:40
A critical evaluation of the optimal association measures for creating L2 learners' collocations lists
Kohei Fukuda（Tokyo University of Foreign Studies, Graduate Student）
Yukio Tono（Tokyo University of Foreign Studies）

研究発表26：14:40-15:00
教科書コーパス分析における推定周辺平均値の有用性について
―英語法助動詞の頻度と意味の分析から―
梶山　達也（同志社大学大学院生）

●研究発表第10セッション　英語学・英文学（場所：ブレイクアウトルーム2）
司会：神原　一帆（京都外国語大学･立命館大学）

研究発表27：14:20-14:40
分布意味論の手法を応用したフレーム意味論の分析
森下　裕三（桃山学院大学）

研究発表28：14:40-15:00
Negated speech and thought presentation in contemporary present-tense fiction
Reiko Ikeo（Senshu University）

●研究発表第11セッション　英語教育・教材（場所：ブレイクアウトルーム3）
司会：宇佐美裕子（東海大学）

研究発表29：14:20-14:40
高校英語教科書におけるコロケーション：基本動詞に着目して
畔元里沙子（九州大学大学院生）

研究発表30：14:40-15:00
Corpus based research on vocabulary development: Focusing on phrasal verbs composed of A-level verbs and particles
Kohei Takebayashi（Tokyo University of Foreign Studies, Graduate Student）
Yukio Tono（Tokyo University of Foreign Studies）

研究発表31：15:00-15:20

A comparative study on collocations used in Japanese junior high school English textbooks and CEFR-based English coursebooks

Noriaki Mikajiri（Tokyo University of Foreign Studies, Graduate Student)

Yukio Tono（Tokyo University of Foreign Studies）

基調講演　15:30-16:20（場所：メインルーム）

What norms for language learners? A corpus-based research and teaching perspective

司会：Dr. Yasuhiro Fujiwara（Meijo University)

講師：Dr. Gaëtanelle Gilquin（The Université Catholique de Louvain, Belgium)

シンポジウム　16:30-17:30（場所：メインルーム）

Introducing VOICE 3.0: ELF perspectives for Learner Corpus Research

司会：Dr. Shin'ichiro Ishikawa（Kobe University)

講師：Dr. Marie-Luise Pitzl-Hagin（Austrian Academy of Sciences, Austria)

閉会式　17:30-17:40（場所：メインルーム）

閉会の辞

田畑　智司（大阪大学）

学生優秀発表賞授賞式

【ワークショップ 1】

Running a Vocabulary Course With Lextutor

Dr. Tom Cobb（Université du Québec à Montréal）

The importance of vocabulary knowledge in any type of language course or curriculum is now acknowledged but still not easy to incorporate in a systematic manner. There are few dedicated vocabulary courses, and the ones there are do not match the increasing specialisation of many learner programs. Lextutor has been designed to basically take on the whole job of a dedicated vocabulary supplement from a practical corpus perspective. My workshop will show the main steps in this process, from building a corpus of learning materials, to placement testing, to text selection and adaptation, to assuring a manageable supply of new and appropriate items, to test writing that reflects what learners have actually been exposed to sufficiently to be tested on. The theme is ‘corpora for courses’ and I will share results from locales where this approach and technology is being deployed.

【ワークショップ 2】

言語データを対象とした KH Coder の活用法

樋口　耕一（立命館大学）

報告者は, 社会調査におけるテキスト分析を当初の目的として,「計量テキスト分析」の方法を提案し, 実践のためのツール「KH Coder」を開発しています。本ワークショップでは，この方法とツールを用いて自由回答（日本語）を分析した研究事例を紹介したうえで，方法の特徴について述べます。また，ご紹介した研究事例で分析したデータを使用し，まったく同じ結果をえるための KH Coder 操作をデモ形式で示します。実際の研究データなので「OK」ボタン押せば完了というわけにはいかず，結果を見ながら微調整していく過程も示します。最後に，英語をはじめとする多言語対応状況についてお話しします。

【ワークショップ 3】

初めての XML

永崎　研宣（一般財団法人人文情報学研究所）

コーパス研究には様々なアプローチがある。近年は自然言語処理技術に頼って超大規模コーパスを処理するという魅力的な手法が広まってきているが，一方で，単語や文章に事前にタグを付けた上でそのタグを利用することで分析を行う手法も未だ健在である。タグを付ける際にはいくつかの手法があり得るが，その一つの手法として XML（Extensible Markup Language）がある。このワークショップでは，XML の概要と応用事例を簡単に紹介した上で，XML タグを活用したコーパスを分析するための基本的な手法について扱う。ツールとしては，Google Colabo で提供される Python を用いることで，参加者の方々が再現・応用しやすいものとなることを目指したい。

【研究発表第 1 セッション】
【研究発表 1】

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

Christopher Cooper（Rikkyo University）

The pre-processing of texts is an important step in any corpus-based research, especially when dealing with internet-based texts, where the data tends to be ‘noisy’. In this presentation, the pre-processing steps taken to prepare YouTube transcripts for a multi-dimensional analysis will be described. Examining the texts and keeping the end goal and purpose of the study in mind is essential when deciding what elements of texts should be edited. For this study, the transcripts had the following characteristics that needed to be cleaned: words that do not represent speech (e.g. ‘[laughter]’), censored words (represented as ‘[__]’), there were no sentence boundaries or punctuation, and all words were lower case (causing tagging problems for proper nouns and lower case ‘i’). Potential solutions to these problems will be discussed including using regular expressions, using a Stanford NLP caseless model to capitalise proper nouns, using an open-source punctuation prediction model that is available as a Python library to add sentence boundaries, and replacing censored words with pseudo words. The accuracy and suitability of the solutions will be reported and any feedback from the audience will be very welcome.

【研究発表2】

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

仁科　恭徳（神戸学院大学）・赤瀬川史朗（Lago NLP）

仁科・赤瀬川（2021，2022）では，現在までに構築された日英・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を網羅的に振り返り，日英・英日パラレルコーパスオンライン検索ツール『パラレルリンク』（Ver. 1.0）に搭載予定であった9種のパラレルコーパスの概要と，それらを再整備する上で施したテキスト処理やアノテーション，全文検索インデックスの作成，ファイル整理などの一連の作業について詳説した。本発表では，当該ツールのプロトタイプの具体的なインターフェースや検索機能，実装されている統計処理，想定される活用紹介などについて紹介する。

【研究発表3】

PEP コーパスプロジェクト : その設計と射程

神原　一帆（京都外国語大学・立命館大学）・木村　修平（立命館大学）・
近藤　雪絵（立命館大学）・山下　美朋（立命館大学）・
山中　　司（立命館大学）

本発表の目的は立命館大学で実施されているプロジェクト型必修英語プログラムにおける学生の term paper を集積した学習者コーパスの設計案とその発展可能性を論じることにある。このプログラムでは自由度の高いプロジェクトが奨励される一方で複数教員の指導による表現や書式，評価の一貫性といった課題がある。本研究はこうした課題を解決するための前準備として位置付けることができる。本発表ではコーパスの設計案を論じ，その発展可能性として，このコーパスが（A）教員によるプロジェクト評価の傾向の調査，（B）ライティングの構造研究，（C）科学論文「らしさ」の測定のための基礎データとして働きうることを主張する。

【研究発表第２セッション】
【研究発表４】

受動文における様態副詞の生起位置に関する一考察

西村　知修（石川工業高等専門学校）

様態副詞は受動文で動詞後方または be 動詞と過去分詞の間（中間位置）に生起できるが，中間位置に生起するのが普通であるとされ（Swan 2016:§201），文成立に必須の副詞は動詞後方に生起できないという指摘もある。

（1）She was badly treated / *treated badly.（安藤 2005: 525）

BNC や COCA で観察すると副詞が受動文で中間位置に高い頻度で現れるのは事実のようだが，動詞と副詞の組み合わせによってその頻度は異なり，be treated badly も散見される。本発表では「副詞＋形容詞（＋名詞）」の結びつきの強弱が受動文における副詞の位置に影響を与えている可能性を探る。

主要参考文献

安藤貞雄（2005）『現代英文法講義』開拓社，東京.
Swan, Michael (2016) Practical English Usage, 4th ed., Oxford University Press, Oxford.

【研究発表５】

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

藏薗　和也（神戸学院大学）

本発表の目的は，起動動詞 start の補部に名詞句（Noun Phrase, NP）がくる表現 start NP にどの様な語句が生起するかに関して一般化を行うとともに，なぜ特定の性質の語句が生起するかについて, 意味的な観点から説明することにある。本調査では, start NP に生起する名詞句を The British National Corpus を利用して抽出し，手作業で質的に分類した。さらに，The Corpus of Contemporary American English を利用して begin/start +NP に生起する名詞句を調査した中村（2018）や類義語の begin NP に生起する名詞句を調査した拙稿（2021）の結果と比較し，さらにネイティブの直観や調査（Freed 1979）を考慮することで，start NP には start の使役の意味から「起動後に自分で動いたり，進展していく性質のもの（a car/engine/a team/company/fire/trouble/game/a war)」が頻繁に生起することを主張の柱にすえて議論を進めていく。

【研究発表6】

学習英文法における as best as possible の位置付け

松田　佑治（関西外国語大学大学院生）

as X as 構文の X には，通常は原形が要求される。しかし，as best as possible や，as less wrong as possible など，X に最上級や比較級が生起している事例が存在する。そこで，本発表では，大規模なコーパス調査に基づき，as X as 構文の X に最上級，比較級が生起する事例を正規表現で抽出し, それぞれの頻度を示す。そして, インフォーマントの意見も踏まえ，その分析結果として，現代英語では，as X as possible の形に限り，最上級 best のみが容認されていることを論証する。その一方，best の反対の worst や比較級にも，次第に拡張されている点をコーパス調査に基づいて説明する。

【研究発表第３セッション】
【研究発表7】

A longitudinal study of fluency based on learner corpus of English conversations

Maxim Tikhonenko（Tokyo University of Foreign Studies, Graduate Student）
Keiko Mochizuki（Tokyo University of Foreign Studies）

The purpose of this study is to analyze the longitudinal growth of fluency in English conversations by three high school learners. The learner corpus consists of video and audio recordings of 30-minute remote one-on-one speaking lesson with a native English-speaking instructor, recorded over a 20-month period from the first year to the third year of high school. In September 2020, three months after the 20th month lesson, all three students took the Aptis speaking test and their scores were 33 out of 50 points, which was judged as Cefr B level.

The method of fluency analysis was as follows. First, speaking data were manually transcribed and analyzed by ELAN software; second, “speech duration and silent pause length” were measured. Third, the dialogue texts were divided into AS-Units based on Foster, P., A. Tonkyn, and G. Wigglesworth (2000), and fluency in the 10th and 20th month lessons of the three learners was analyzed based on the following criterions.

1) Speech rate (words/total time)　　2) Pause rate (total pause time/speech time)

The results of the speech rate analysis show that all three learners spoke more words per minute and the number of words per minute increased at the 20th month.

Reference

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.

【研究発表 8】

日本人中学生英語学習者の take の使用と語義指導への提言

澤口　　遼（関西大学第一中学校）

基本語 take の語義の多くは中学校で導入されるが，学習者の習熟度に応じてどの段階で各語義を提示・指導すべきかは十分に検討されていない。そこで本研究では，英語学習者コーパス JEFLL の中学生学習者が使用する take の語義を CEFR レベル別語義データベース English Vocabulary Profile によって分類し，各学年でどの CEFR レベルの語義が使用可能になるのか調査した。その結果，学習者の使用する take の語義はほとんどが A1～B1 レベルで，A2 レベルの語義の使用が少ないこと，学年が上がるにつれて B1 レベルの語義の使用が増加することが判明した。教科書等の教材編集への示唆として，A2 レベルの語義の記述を充実させる，学年を考慮して B1 レベルの語義を導入することが挙げられる。

主要参考文献

投野由紀夫（編）（2007）『日本人中高生一万人のコーパス : JEFLL Corpus』小学館.

Capel, A. (2015). The English vocabulary profile. In J, Harrison. & F, Barker. (Eds.), English profile studies 5, pp. 9–27. Cambridge: Cambridge University Press.

【研究発表 9】

学習者コーパスの品詞連鎖分析による母語の英語構文への影響

山口　一華（東京外国語大学大学院生）
投野由紀夫（東京外国語大学）

本研究は，第二言語習得における母語の英語構文への影響を調査するため，母語の異なる英語学習者の言語データを収録した学習者コーパス中の品詞連鎖を分析した。特に本研究では，ICCI コーパスのオーストリアサブコーパスと JEFLL コーパスを用いた。ドイツ語母語話者と日本語母語話者の初級～中級 EFL 英語学習者データに出

現する単語を品詞タグに変換し，n-gram形式で品詞連鎖を取り出し，その傾向を調べた。さらに，抽出されたn-gram統計をもとに，英語母語話者コーパスで得たn-gram統計を基準に，2つの異なる母語の学習者データを比較し，ドイツ語母語話者と日本語母語話者それぞれに特有の品詞連鎖の過剰・過少使用を明らかにする。

【研究発表第4セッション】
【研究発表10】

研究論文における英語と日本語のアブストラクトは読み手を意識しているか

浅野　元子（大阪医科薬科大学）・松田　紀子（近畿大学）

本研究は，論文における英語と日本語の抄録が各々の読み手を意識した言語的特徴を有するかどうかを探索した。2020年1年分の教育心理学研究誌（EP）30報と日本公衆衛生雑誌（JPH）30報の英日抄録テクストをコーパス化し，修辞の運びを分析した。EPとJPHの総語数は英文が約6,000語と14,000語(CasualConc), 日本文は約9,000語と16,000語（KH Coder）であった。双方の約半数は英文と日本文の論理の運びに差異があり，EPは英文で研究の対象，データ収集方法や手順，JPHは様々な部分の英文または日本文で詳述しており，各々の読み手を意識して調整されていることが示唆された。

【研究発表11】

高校生のライティングに対する添削者の添削エラー分析
—時制の誤りに焦点を当てて—

竹下　綾音（九州大学大学生）

本研究は，日本の高校生の英作文と添削者の英文を用いてパラレルコーパスを作成し, 高校生の中間言語における時制に関する誤用と, それに対する添削者の添削エラーの傾向を分析することを目的とする。これまで学習者の中間言語に関する研究は多く行われてきたが，ライティングに対する添削やその誤りに関する分析はほとんど行われていない。本研究では, 日本語を母語としない添削者（フィリピン, バングラデシュ等在住者）による添削データを対象とし，時制に関する添削誤りを中心に考察する。分析の結果，日本語の表現（～ている等）に引きずられた高校生の誤りについて特に添削エラーが多いことが明らかになった。

【研究発表 12】

Development of the Ballet English Corpus (Ver. 1.0) for art major students

Hiroko Usami（Tokai University）

English for specific fields such as business, law, science technology, medicine, nursing, tourism, aviation has been examined, and specialised corpora for specific English fields have been constructed and applied to teaching. However, English used in the field of art, including music, performing arts, fine arts, and dance, especially classical ballet, has been insufficiently examined.

Various types of dance, including, in particular, classical ballet, have been enjoyed by people of all ages for different purposes in Japan. An increasing number of young Japanese dancers are learning classical ballet both abroad and in Japan and are dancing in overseas ballet companies.

Therefore, this study aims to introduce a specialised corpus, the Ballet English Corpus (BEC), which can be applied to the study of English for specific purposes. The BEC (Ver. 1.0) contains written texts in ten different categories related to classical ballet (ballet techniques, companies, studios, history, schools, theatres, people, narratives, repertoires, and miscellaneous) that are used by both adults and children. This study describes the design and analysis of the word lists in the BEC.

reference

Showa Academia Musicae. (2022). Nihon no ballet kyoiku ni kansuru zenkoku chosa - hokokusyo. [Nationwide survey on ballet education in Japan – Report]. Showa Academia Musicae.

【研究発表第 5 セッション】
【研究発表 13】

COCA における正誤表現と非標準表現について

田畑　圭介（神戸親和女子大学）

誤用と区別される非標準用法や正用法の判定基準について，COCA で検出される頻度情報をもとに考察を行う。検証対象となる英語表現は What do you got?/get me to doing/got A confused with someone else 等である。帰結 1：COCA において正用法との比率が 15％を超える用法で TV/Movie で一定の使用が確認できる用法は非標準用法と

してその存在が認定できる。帰結 2：COCA において 200 以上の用例が確認され，正用法との対立がない表現，あるいは正用法と同程度の頻度と生産性が確認できる用法は，標準用法 (= 正用法) と認定できる。

【研究発表 14】

CasualConc 3.0 - Universal Dependency タグを利用した文法検索の試み

今尾　康裕（大阪大学）

macOS 用アプリケーション CasualConc は，基本的なコーパス分析ツールが備えている KWIC 検索，単語リスト作成，コロケーション検索などの機能に加えて，統計環境 R を利用したグラフ作成機能を備える統合的コーパス分析環境である。今春公開した新しいバージョンでは，XML タグを利用した簡易的な検索機能やファイルのフィルタリング，Stanford CoreNLP による Universal Dependency タグを利用した文法検索などの機能を追加した。本発表では，新機能を紹介するとともに開発の課題なども議論する。

主要参考文献

Imao, Y. (2002). CasualConc (Version 3.0.3) [Computer software].

【研究発表 15】

量的概念分析再考 : 動詞 explain を例に

菅原　裕輝（大阪大学）・神原　一帆（京都外国語大学・立命館大学）

近年，科学哲学分野において，内省に基づく分析に対する方法論批判を受け［1］，量的手法を組み込んだ概念分析の開発が行われてきている［2］。本研究では，量的な科学哲学の方法論の可能性と限界について，特に共起語情報のみに基づく量的な手法を批判的に検討し, 量的な認知意味論の方法を組み入れた量的科学哲学の方法を提案･実行する。この目的の下，BNC から無作為に抽出した 657 件の動詞 explain を含む事例に対し，フレーム意味論的な解析に加え［3］，様々な構文解析［4］等を行った。この結果，動詞 explain が喚起する概念構造をより緻密に表すだけでなく，量的分析によって先行研究での結果の比較が可能となった。

［1］Overton, J. A. (2013). “Explain” in scientific discourse. Synthese, 190(8), 1383–1405.
［2］Pence, C. H., & Ramsey, G. (2018). How to do digital philosophy of science. Philosophy

of Science, 85(5), 930-941.

［3］Fillmore, C. J., Johnson, C. R., & Petruck, M. R. L. (2003). Background to FrameNet. International Journal of Lexicography, 16(3), 235–250.

［4］Gries, S. T. (2010). Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. The Mental Lexicon, 5(3), 323–346.

【研究発表第６セッション】

【研究発表 16】

米国一般教書演説に出現する分離不定詞の効果に関する予備的研究

福本　広光（大阪大学大学院生）

本研究は，歴代米国一般教書演説（State of the Union Address）のスクリプトをコーパスとして，そこに現れる分離不定詞（split infinitive, 下記例文参照）の効果および使用意図の一端を探ろうとするものである。

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

The American Presidency Project 所収の演説原稿に品詞タグを付与し，AntConc でコンコーダンス検索を行って用例を抽出した。分離不定詞は時代や人物ごとの揺れはあるものの米国一般教書演説においても一定数用いられており，本発表では，“splitter”（to 不定詞を「割る」副詞相当語句）を軸として，収集した全用例を分類し，大統領が分離不定詞を使うことで強調していると考えられる事柄について，共起表現にも気を配りつつ，幾つかのコンテクストを分析する。

【研究発表 17】

男性リーダーの発話データの特徴分析：
オックスフォード・ユニオン及び TED Talk 分析の示唆

中谷　安男（法政大学）

今や性別によるリーダーの役割の差をつけるべきではない。この点に配慮しつつ，本発表は，ビジネス・政治・社会活動に関わる男性リーダーの英語発話における語彙的特徴を調査した。世界のリーダーを多く輩出しているオックスフォード・ユニオンにおけるディベートやスピーチのコーパスデータを作成した。これに加えて，TED

Talkのリーダーのコーパスを作成した。合計164人のデータをAntConc 4.0.10を活用し，Keyword分析により，男性リーダーの特徴語を抽出した。結果としてyou, get, go, this等が特徴語彙として抽出された。本報告では，これらのクラスター分析も活用し特徴的な表現の具体例を示す。

【研究発表18】

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

島津美和子（立教大学）

現在，日本政府が進める改憲の理由の一つに，現行憲法はGHQから提示された英語原案を日本語に翻訳したものであり，日本が策定したものではないという主張がある。この主張の言語的妥当性を検証するため，本研究では，帝国議会・国会議事録をコーパスとみなし，憲法の文脈における「翻訳」の用例を抽出し，分析した。その結果，「翻訳」は「翻訳調」「翻訳臭（い）」といった特定の句で用い，否定的なニュアンスを伴うことが分かった。一方，米国議会議事録からtranslate/translationとConstitutionが共起する事例を抽出し，分析した結果，合衆国憲法の他言語訳は肯定的に受け止められていることが分かった。

【研究発表第7セッション】
【研究発表19】

日本人英語学習者の「うなぎ文」の使用に関する分析

藤原　康弘（名城大学）・岩男　考哲（神戸市外国語大学）・
伊藤　　創（関西国際大学）・仲　　　潔（岐阜大学）

本発表の主たる目的は，いわゆる「うなぎ文」が，日本人英語学習者による英語に確認されるか，また「英語力」の伸長と共にどのような変化がみられるかをコーパス言語学的手法で検証することである。日本語は名詞／形容詞述語文に「僕は珈琲だ」「今日は忙しい」のように解釈においてコンテクストに依存する部分が大きい「うなぎ文」と呼ばれる表現がある。第二言語習得や英語教育の先行研究では，日本人英語学習者の英語に一定程度確認されることが指摘されてきたが，能力の伸長による変化の詳細はまだ明らかではない。本研究では，International Corpus Network of Asian Learners of English（Ishikawa, 2013）を用いて，うなぎ文の一部の使用傾向の分析結果を報告し，広く議論を行う。

【研究発表20】

The use of nominalization features in the academic texts by Japanese learners

Kaede Hanawa（Tokyo University of Foreign Studies, Graduate Student）
Yukio Tono（Tokyo University of Foreign Studies）

Since the introduction of the term "grammatical metaphor" by Halliday (1985), nominalization features have caught attention in the research of academic written texts (Biber & Gray, 2013). The present study aims to provide a descriptive view of nominalization features found in the academic writing produced by Japanese L2 learners of English in comparison with L1 users of English.

Corpora used in this study are ICLE-JP for Japanese learners and LOCKNESS for L1 users. Sketch Engine is used for the collocation search. Using the CQL function, all the sentences which include "noun + preposition" are selected for the further annotation phase. Those items are then annotated and categorized under several types of nominalizations. The two sets of data are compared in terms of the type frequency and the verbs (both intransitive and transitive verbs) or adjectives which the nominalizations originate from.

References

Biber, D., & Gray, B. (2013). The Verb Phrase in English: Nominalizing the verb phrase in academic science writing.

Halliday, M. A. K. (1985). An Introduction to Functional Grammar. London: Edward Arnold.

【研究発表21】

The misuses of English articles in compositions of L1 Chinese and Japanese learners: A corpus-based study

Xiao Sun（Kyoto Sangyo University, Graduate Student）

English articles belong to a category of high frequency words. But research on the article system has been mainly focused on the analysis of functions and description of usage, with few studies focusing on the distribution and features of English article errors (Zhou, Xia, Du & Yan-xia, 2015). At the same time, corpus-based research on English articles has also been limited. Hence, this study analyzes misuses of English articles made by L1 Chinese and Japanese learners of English using written corpora. Data is collected from the Nagoya Interlanguage Corpus of English Reborn (NICER) and the Chinese Learner English Corpus (CLEC). In these corpora a wide range of variability in misuse can be found. As 和泉 et. al (2004) have pointed out, the misuses of English articles can be categorized into three patterns: omission of articles, substitution of articles and over-employment of articles, but it still can be predicted that there are some differences existing between these groups of learners in their use of English articles. It is hoped that the findings of this comparative study can have a positive effect on the article acquisition education for L1 Chinese learners and L1 Japanese learners.

主要参考文献

Zhou, Xia, Du, & Yan-xia (2015). An Investigation of the Misuse of English Articles of Chinese English Learners.

和泉絵美･内元清貴･井佐原均(2004). 日本人英語学習者の英語冠詞習得傾向の分析. 『日本人 1200 人の英語スピーキングコーパス』東京：アルク

【研究発表第 8 セッション】

【研究発表 22】

初等・中等教育向け DDL ツールが目指す「自律的なコーパスユーザーの育成」

西垣知佳子（千葉大学）・赤瀬川史朗（Lago NLP）

発表者らは小学生向けと中高生向けの「英語用例コーパス」と「DDL ツール」をそれぞれ開発し，初等・中等教育の英語授業における DDL（data-driven learning）の普及を目指し，学校現場で実践を行っている。発表者らの DDL 開発では，短期･中期･長期の 3 つの目標を設定している。最終的な到達点となる長期目標には「小学生の頃からコンコーダンスラインに慣れ親しんだ自律したコーパスユーザーの育成」を掲げている。本発表ではこの 10 年間にわたる DDL ツールの開発を振り返り，長期目標の「コンコーダンスラインに慣れ親しむ」ことを目指した中高生向けの DDL ツールの新たな機能について報告する。

【研究発表 23】

中学校通常英語授業における DDL の活用を目指して：
学習環境・タイミングの違いによる比較

中井　康平（千葉大学教育学部附属中学校）・水本　　篤（関西大学）・
西垣知佳子（千葉大学）

本研究では，中学 1 年，2 年，3 年の約 260 名が，自主開発した中高生用 DDL 支

援ツールを使って，自宅と学校の異なる学習環境で，また，予習，復習，あるいは教科書学習の補助という3種のタイミングで英文法を学習した後，質問紙に回答した。質問紙には5件法と自由筆記の調査があった。分析の結果，DDLの使用環境について自宅と学校では勉強のしやすさ等の点で差が見られなかった。また，発見活動にかかる時間の節約を目標にして行った教科書学習の補助のDDLが，生徒のわかりやすさにつながるという結果が得られた。これらの結果から，中学校の通常英語授業に恒常的にDDLを導入する場合の可能性，そして問題点を議論する。

【研究発表24】

英作文における動詞—名詞コロケーション産出に対するDDLの効果

佐竹　由帆（青山学院大学）

コロケーションの知識は重要だが，英語学習者にとってコロケーションを正確に使用することは難しい。筆者の先行研究（佐竹2021）はコーパスを参照して学習するデータ駆動型学習（DDL）が動詞・名詞コロケーションを覚える上で有効であることを示唆したため，本研究は英作文におけるDDLの動詞–名詞コロケーション産出に対する効果を検証した。被験者は20名の日本の大学1年生で英語中級学習者であり，筆者が選んだ2つの動詞–名詞コロケーションをWordbanks Onlineで検索して用例を見る学習を毎週1度9週間行った。被験者が学習の前後に書いた英作文を比較した結果，事前英作文ではほとんど使用されなかった対象コロケーションが事後の英作文では使用されており，コロケーション産出に対するDDLの有効性が示唆された。

【研究発表第9セッション】
【研究発表25】

A critical evaluation of the optimal association measures for creating L2 learners' collocations lists

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

Association measures (AM) have been used to extract candidates for the collocations list for L2 learners. However, little attention has been paid to which AM is optimal for extracting collocations for pedagogical purposes. Furthermore, dispersion measures are often not considered for the evaluation of AMs. This study explores which AM is suitable for identifying pedagogically useful English collocations, considering both frequency and dispersion thresh-

olds. Verb-Object and Modifier-Noun pairs were extracted from the syntactically parsed British National Corpus, and their AMs (MI, MI3, T-score, Z-score, logDice, Log-likelihood) were calculated. Dispersion was calculated using Gries' DP. AMs were evaluated against the gold standard, defined as the items found in CEFR-based English coursebooks and Oxford Collocations Dictionary (OCD). Precision and recall were evaluated by setting varying frequency and DP thresholds. The main results are as follows: (a) without DP threshold, T-score is the most closely associated with the selection of collocations in the gold standard; (b) moderate DP threshold can enhance the performance of AMs, (c) in most cases, DP threshold ≦ 0.4 has an adverse effect on the performance of AMs; and (d) frequency thresholds cannot enhance the performance of Log-likelihood, but DP threshold can make it surpass T-score in some cases.

【研究発表 26】

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

梶山　達也（同志社大学大学院生）

ここでは, 教科書コーパス分析に対する推定周辺平均値の有用性について発表する。現在, 大規模コーパス分析にはカイ二乗検定がよく使われている。しかし, 教科書コーパスのような規模の小さいコーパスに対して，大規模コーパス分析で使用されるカイ二乗検定を用いると，正しく検定されない可能性が指摘されている。そこで，2022年3月開催英語コーパス学会語彙研究会で発表した，英語助動詞の頻度と意味に関する中・高英語教科書コーパスの分析結果を批判的に考察しながら，新たに，推定周辺平均値を用いた分析を行い，これの有用性について議論する。

【研究発表第 10 セッション】
【研究発表 27】

分布意味論の手法を応用したフレーム意味論の分析

森下　裕三（桃山学院大学）

本研究では，land と ground という名詞，および stingy, thrifty, generous, wasteful という形容詞を対象として認知意味論が想定する意味観に基づいた分析（e.g. Fillmore 1982）をおこなう。本研究では，COCA において，上記の語が生起する全用例を対象として分布意味論（e.g. Montes and Heylen 2022）の手法に基づく分析をおこなった。結果として，認知言語学で主張されてきた land と ground という語の対称性が確認さ

れたものの，stingy と thrifty という語の関係，ならびに generous と wasteful という語の関係については，先行研究において主張されていた以上に複雑な反義関係・類義関係にあることを示す結果を得た。

主要参考文献

Fillmore, Charles J. (1982) Frame semantics. In The Linguistic Society of Korea (ed.), Linguistics in Morning Calm. 111–137. Seoul: Hanshin Publishing Company.

Montes, Mariana and Kris Heylen (2022) Visualizing Distributional Semantics. In Dennis Tay & Molly Xie Pan (eds.), Data Analytics in Cognitive Linguistics. Methods and Insights. 103–136. Berlin: Mouton De Gruyter.

【研究発表 28】

Negated speech and thought presentation in contemporary present-tense fiction

Reiko Ikeo（Senshu University）

This paper shows how negated speech and thought presentation functions in narrative by using a corpus approach. A corpus consisting of texts from contemporary present-tense fiction was annotated with discourse presentation categories based on the Semino and Short model (2004). In this annotation process, a new subcategory ‘g’ has been introduced for negated discourse presentation. The data shows that negated thought presentation occurs more frequently than negated speech presentation. The subcategory “g” is attached to 29 out of 3,261 speech presentation cases, which accounts for 0.9% of all the speech presentation tags. In contrast, this subcategory is found in 150 cases out of 1,805 thought presentation cases, which accounts for 8.3% of thought presentation tags.

Negated cases of speech and thought presentation reveal characters’ inner worlds in a way that affirmatives do not. In contrast to events and states which are expressed by means of affirmative terms, “non-events” and “non-states” which are expressed by means of negatives are usually less salient and less informative. When non-events or non-states are expressed in narrative, they stand out and thus have special textual effects.

Reference

Semino, E. and Short, M. (2004) Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing. London: Routledge.

【研究発表第 11 セッション】

【研究発表 29】

高校英語教科書におけるコロケーション：基本動詞に着目して

畔元里沙子（九州大学大学院生）

本研究は，高校英語教科書で扱う基本動詞を含むコロケーションをリスト化し，一般コーパスにおけるコロケーションと比較することを目的とする。具体的には，高校英語教科書コーパスから基本動詞を含むコロケーションを抽出し，一般コーパスでの T スコア・頻度等を付与して重要度・出現率等を評価する。さらに，教科書コーパスを学年別に分類し，学年ごとのコロケーションの相違や，一般コーパスにおけるコロケーションカバー率の変化を分析する。

【研究発表 30】

Corpus based research on vocabulary development: Focusing on phrasal verbs composed of A-level verbs and particles

Kohei Takebayashi（Tokyo University of Foreign Studies, Graduate Student）
Yukio Tono（Tokyo University of Foreign Studies）

Despite the significance of phrasal verbs (PVs) in communication, learners have a noticeable tendency to avoid PVs in favour of one-word equivalents. The author argues that PVs would serve as significant building blocks to develop learners' vocabulary knowledge. To this end, this study explores the possibility of using PVs as a bridge between A-level verbs and B-level verbs as defined by the CEFR. A corpus of PV textbooks (size = 1.2 million) was compiled, and pairs of PVs and their single-word verb (SV) equivalents were retrieved. After producing a list of [PV – SV] pairs, the vocabulary levels of those verbs on the list were identified in accordance with the English Vocabulary Profile to investigate the extent to which PVs can be replaced with their SV counterparts. At the same time, their semantic opacity and the degree of semantic transformation between PVs and their SV equivalents were examined. The results show that PVs have the potential to serve as a bridge between A-level and B-level verbs, and a selected group of PVs will make a significant impact on the expansion of the range of meaning related to verb semantics.

【研究発表31】

A comparative study on collocations used in Japanese junior high school English textbooks and CEFR-based English coursebooks

Noriaki Mikajiri（Tokyo University of Foreign Studies, Graduate Student）
Yukio Tono（Tokyo University of Foreign Studies）

Collocations are arguably one of the most important points for learning a language. The present study investigates the use of English collocations by comparing authorized English textbooks published in Japan and English native corpora in order to improve the contents of English textbooks. To this end, three corpora were used in this study: a corpus of English textbooks for junior high school students, a corpus of CEFR-based English coursebooks published in the UK, and the British National Corpus (BNC) as a reference corpus. From each textbook corpus, a list of verb-object collocations was extracted for each CEFR level via Sketch Engine. For each extracted collocation list, frequency and association measures such as t-score, MI score, logDice, Log-likelihood were obtained using the BNC. Based on those results, extracted collocations used in one corpus but not the other were compared. I also analyzed how the collocation lists would differ when evaluated by each association measure. The identification of collocations and association measures that are markedly different from the collocations used in the textbook and coursebook corpora will allow us to find other collocations to learn from the BNC and other native speaker corpora and make suggestions for textbook improvement.

【講演】

What norms for language learners? A corpus-based research and teaching perspective

Dr. Gaëtanelle Gilquin（The Université Catholique de Louvain, Belgium）

In both research and teaching, norms are often used as a reference against which to describe or evaluate learner language (L2). In learner corpus research, for example, the notions of ‘underuse’ and ‘overuse’ imply a comparison of frequencies in the L2 and in some variety corresponding to the expected target (see Granger 1996). In language teaching, textbooks mostly provide materials representing native varieties of the language and teachers often assess learners’ proficiency with reference to some native standard. Corpora, by giving access to data that are representative of a certain language or language variety, can serve as a basis to define a norm empirically (Klippel & Mukherjee 2007).

This presentation will provide an overview of the different types of corpus-derived norms that can be used in learner corpus research and foreign language teaching. We will consider whether a norm is always required and will examine the use of native vs non-native norms, novice vs expert norms, single vs multiple norms and research vs pedagogical norms, among others. It will also be shown that using different norms can lead to different results (see Chen 2013 for an example), which means that it is crucial to choose the most appropriate norm(s). In this respect, the importance of context will be underlined, including one's research purposes (cf. Ädel 2006) and learners' goals in studying the language.

It will be argued that norms are useful in most L2 research and teaching contexts but should be carefully chosen to reflect the most relevant target languages/varieties and should be applied with some flexibility to allow, for example, for linguistic creativity.

References

Ädel, A. (2006). *The Use of Metadiscourse in Argumentative Texts by Advanced Learners and Native Speakers of English*. Amsterdam: John Benjamins.

Chen, M. (2013). Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics* 18(3): 418–442.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (eds) *Languages in Contrast. Text-based Cross-linguistic Studies* (pp. 37–51). Lund: Lund University Press.

Klippel, F. & Mukherjee, J. (2007). Standards and norms in language description and language teaching: An introduction. In S. Volk-Birke & J. Lippert (eds) *Anglistentag 2006 Halle, Proceedings* (pp. 303–306). Trier: Wissenschaftlicher Verlag Trier.

【シンポジウム】

Introducing VOICE 3.0: ELF perspectives for Learner Corpus Research

Dr. Marie-Luise Pitzl-Hagin（Austrian Academy of Sciences, Austria）

Since the first online release of the Vienna-Oxford International Corpus of English (VOICE) in 2009, thousands of users world-wide have used VOICE for linguistic research and university teaching to explore the nature spoken English as a lingua franca (ELF) interactions. This talk introduces the new version of VOICE – VOICE 3.0 – built in the VOICE CLARIAH project (2020-2021) and released in September 2021. It provides a tour of the new open-access interface VOICE 3.0 Online that was developed by an interdisciplinary team at the Austrian

Centre for Digital Humanities and Cultural Heritage (Austrian Academy of Science) and the University of Vienna. During the talk, we will have a look at key features of VOICE 3.0 Online like its enhanced filter, style and search functions. In doing so, we will consider the expanded research potential of VOICE 3.0. New technical features include for instance being able to search for select spoken mark-up (such as pauses and overlaps) and being able to combine lexical and mark-up queries flexibly with queries for part-of-speech (POS) tags. The talk will also offer a brief glimpse at the technology used to create the new frontend and backend infrastructure for VOICE 3.0.

Following the rationale of VOICE as an ELF corpus, the second part of my talk will then propose and map out some ELF perspectives for Learner Corpus Research (LCR). Here, we will specifically concern ourselves with the issue of corpus annotation and highlight differences in approach and practice between ELF research and LCR. In order to make tangible these differences, the talk will, among other things, report on an ongoing research project (see Riegler in press) that develops a mark-up scheme for annotating pragmatic functions in VOICE. A major difference between an ELF approach and an LCR approach to corpus annotation lies in the way the two fields address the issue of norms and normativity. While LCR continues to orient quite strongly to the norms of the so-called target language (e.g. error tagging), ELF research promotes a very different orientation to linguistic norms. From an ELF perspective, norms are always flexible and seen as situationally negotiated and negotiable by participants in a particular interaction. This orientation in ELF research has clear implications for developing pragmatic corpus annotation and, as a next step, direct consequences for teaching implications and applications that can be drawn developed on the basis of such annotation.

References

Riegler, Stefanie. Accepted/forthcoming. Annotating VOICE for pedagogic purposes: the case for a mark-up scheme of pragmatic functions in ELF interactions. In Harrington, Kieran & Patricia Ronan. Demystifying corpus linguistics for English language teaching. London: Palgrave Macmillan.

VOICE. 2021. The Vienna-Oxford International Corpus of English (version VOICE 3.0 Online). Founding director: Barbara Seidlhofer; Principal investigators VOICE 3.0: Marie-Luise Pitzl, Daniel Schopper; Researchers: Angelika Breiteneder, Hans-Christian Breuer, Nora Dorn, Theresa Klimpfinger, Stefan Majewski, Ruth Osimk-Teasdale, Hannes Pirker, Marie-Luise Pitzl, Michael Radeka, Stefanie Riegler, Barbara Seidlhofer, Omar Siam, Daniel Stoxreiter. https://voice3.acdh.oeaw.ac.at (last accessed 24 August 2022).

『英語コーパス研究』投稿規定

（2023年５月改定）

1．投稿資格

投稿は会員に限る。共著の場合，第一著者は会員であることとし，その他の共著者については会員でなくてもよい。

2．原稿の種類と長さ

【研究論文】

英文　A4サイズ　１ページあたり35行（文字数の指定はしない），周囲の余白１インチ（25.4mm），投稿時17ページ以内（Times New Roman 10.5ポイント使用）

和文　A4サイズ　１ページあたり35行（文字数の指定はしない），投稿時17枚以内（明朝体フォント（游明朝･ヒラギノ明朝など）10.5ポイント使用）

※和文中の英文のフォントについては Times New Roman を原則とする。Century は用いてはならない。

いずれも Abstract（英文300語以内），図表，注，参考文献目録，付録，謝辞，著者情報などを含む。

【研究ノート，総説論文・書評論文（Review article, Book review）】

・研究ノート：論文のカテゴリーに属さない小論文や萌芽的な研究，新しい研究開発の成果などをまとめたもの

・総説論文：体系的かつ網羅的に先行研究をまとめたもの

・書評論文：専門書の研究分野への貢献と課題点を明確にしたもの

英文　A4サイズ　１ページあたり35行（文字数の指定はしない），周囲の余白１インチ（25.4mm），投稿時12ページ以内（Times New Roman 10.5ポイント使用）

和文　A4サイズ　１ページあたり35行，投稿時12枚以内（明朝体フォント（游明朝・ヒラギノ明朝など）10.5ポイント使用）

※和文中の英文のフォントについては Times New Roman を原則とする。Century は用いてはならない。

いずれも Abstract（英文300語以内），図表，注，参考文献目録，付録，謝辞，著者情報などを含む。

【その他（ソフトウェアレビュー，書評（図書紹介），コーパス紹介など）】

研究論文の半分以内の分量

3. 原稿作成時の注意

下記のように投稿者を特定できるような情報，その他，本人の同定につながると考えられる情報は，採用決定後の最終原稿に追記するものとし，投稿時には投稿原稿には記載しないこと。

（1）謝辞など

（2）「本論は，英語コーパス学会第X回大会において口頭発表した内容に加筆修正を施したものである。」などの文言

（3）「筆者が収集し，https://... で公開しているデータ…」など，筆者特定につながるURL情報など

（4）本文中で投稿者自身の研究を言及する場合,「XXX（2006）で論じたように…」などと記して，参考文献には当該文献を掲載しないこと。

4. 提出方法など

（1）下記の（A）原稿ファイル（Microsoft Wordで作成したファイルとそのPDFファイル），（B）著者情報ファイル，（C）論文投稿チェックシートの3種類のファイルを電子メール添付で提出。（B），（C）についてはWeb掲載のフォーマットを使用のこと。

（2）電子メールの件名（Subject）は「『英語コーパス研究』投稿原稿（著者氏名）」とすること。

（3）提出先，締め切り期日等に関しては学会Webサイトを参照のこと。

（A）原稿ファイル

a. 提出するファイル名は「原稿題名（著者氏名）」とすること。

b. 原稿題名の前に「論文」,「研究ノート」,「総説論文」,「書評論文」,「コーパス紹介」などの種類を明記すること。

c. 原稿本体の冒頭には上記種類の別と題名のみを記すこと。

（B）著者情報ファイル：「著者情報（著者氏名）」

a. 和文原稿の場合は英文タイトル，英文原稿には和文タイトル

b. 著者氏名（ふりがな・ローマ字表記）

c. 所属

d. 郵便番号・住所・電話番号

e. 電子メールアドレス

（C）論文投稿チェックシート：「論文投稿チェックシート（著者氏名）」

Web掲載のチェックシートの必要項目すべてに☑を入れること。

5. スタイル

投稿論文は，研究論文，研究ノート，総説論文・書評論文の別，また，和文・英文の別にかかわらず，APA（American Psychological Association）Style の最新版，および「論文投稿チェックシート」に従い執筆することとする。

6. 掲載論文等の電子化

掲載された論文等の著作者は，論文等を電子化して学会ホームページで公開することに同意する。

7. 著作権

掲載された論文等の著作権は, 本学会に帰属する。本学会は掲載論文等を印刷媒体・電子媒体で公開する権利を有するものとする。ただし，著作者が自著論文等を自分のホームページに掲載したり，自著の本に転載したりすることは妨げない。

8. 研究倫理

（1）投稿にあたっては，下記文書などを参照し，不正行為のないようにすること。独立行政法人科学技術振興機構『研究者のみなさまへ～研究活動における不正行為の防止について～』https://www.jst.go.jp/contract/kisoken/h25/others/h25s805others131120.pdf

（2）ChatGPT のような大規模言語モデル，および類似の AI ツールは，本文校正，資料・文献リストの整理・確認等，他者の助力を受けても著作者の意匠の範囲を超えないと従来みなされてきた使用範囲内にとどめること。

英語コーパス学会会則

（名称）

第１条　本会は「英語コーパス学会」（Japan Association for English Corpus Studies，略称 JAECS）と称する。

（目的）

第２条　本会は英語コーパス及びコーパスツールの開発・評価・利用に関わる研究，また，英語コーパスを用いた言語研究・言語教育研究・関連研究を促進することを目的とする。

（事業）

第３条　本会は前条の目的を達成するために，次の事業を行う。

(1)　大会・研究会等の開催

(2)　学会誌・学会報等の発行

(3)　その他本会の趣旨に沿う事業

（会員）

第４条　本会の会員は一般会員，学生会員，団体会員，賛助会員，功労会員及び名誉会員よりなる。

(1)　一般会員は本会の趣旨に賛同する個人とする。

(2)　学生会員は本会の趣旨に賛同する個人のうち，大学又は大学院に籍を置く学生とする。

(3)　団体会員は本会の趣旨に賛同する大学，研究所，図書館その他の研究・教育団体とする。

(4)　賛助会員は本会の趣旨に賛同する企業等とする。

(5)　功労会員は本会の活動に長く寄与した個人とする。功労会員の規程は別に定める。

(6)　名誉会員は本会の活動に特別に寄与した個人とする。

（会費）

第５条　本会の会費について以下の通り定める。

(1)　会員は所定の会費を納めるものとする。

(2)　会費の額については次の通りとする。

一般会員	年額	5,000円（在外会員は年額6,000円）
学生会員	年額	2,000円（在外会員は年額3,000円）
団体会員	年額	5,000円
賛助会員	年額	15,000円

(3)　会費は入会時点又は会計年度開始時点で納入する。

(4)　２年間にわたって会費納入がない場合は会員の資格を失う。

(5) 名誉会員，功労会員，顧問からは会費を徴収しない。

（会計年度）

第６条　本会の会計年度は４月１日に始まり，翌年３月31日をもって終わる。

（組織）

第７条　本会に執行部，事務局，役員会，学会誌編集委員会，学会賞選考委員会，大会実行委員会，研究会（SIG）を置く。

(1) 執行部は会長，副会長，事務局長，事務局員で構成し，本会全体にかかわる事業を執行・監督する。

(2) 事務局は事務局長及び事務局員で構成し，本会の事務を執行する。

(3) 役員会は役員で構成し，本会にかかる諸問題を審議・決定する。

(4) 学会誌編集委員会は学会誌の刊行にかかる業務を担当する。学会誌編集委員会の規程は別に定める。

(5) 学会賞選考委員会は学会賞・奨励賞の選考にかかる業務を担当する。学会賞選考委員会の規程は別に定める。

(6) 大会実行委員会は大会の企画・準備・実施にかかる業務を担当する。大会実行委員会の規程は別に定める。

(7) 研究会（SIG）は会員のうち，希望する者によって構成し，それぞれが掲げる研究目的に応じた活動を行う。研究会の規程は別に定める。

（役員）

第８条　本会に次の役員をおく。

(1) 会　長　１名

(2) 副会長　若干名

(3) 理　事　若干名

(4) 幹　事　若干名

（役員の任期・定年）

第９条　役員の任期は以下の通りとする。

(1) 会長・副会長の任期は２年とし，引き続き２期までの再任を妨げない。

(2) 理事・幹事の任期は２年とし，再任を妨げない。

(3) 任期は当該年度の４月１日から起算する。

(4) 役員の定年を70歳とする。任期の途中で定年に達したときは当該年度の終了まで，その任にあたる。

（役員の任務）

第10条　役員の任務は以下の通りとする。

(1) 会長は本会を代表し，会務を統括する。会長は総会・役員会を招集し，これを主宰する。

（2）副会長は会長の命ずる職務を所掌するとともに，会長を補佐し，必要に応じて会長の職務を代行する。

（3）理事は役員会に出席し，本会の運営に関わる重要事項を審議･議決する。

（4）幹事は役員会に出席し，理事を補佐し，本会の運営に関わる重要事項を審議・議決する。

（役員の選出）

第11条　役員は役員会における投票によって決定する。

（役職員）

第12条　本会に次の役職員をおく。

（1）顧問	若干名
（2）事務局長	１名
（3）事務局員	若干名
（4）監査	１名
（5）学会誌編集委員会委員長	１名
（6）学会誌編集委員	若干名
（7）学会賞選考委員会委員長	１名
（8）学会賞選考委員	若干名
（9）大会実行委員会委員長	１名
（10）大会実行委員	若干名

（役職員の任期・定年）

第13条　役職員の任期は以下の通りとする。

（1）顧問の任期は終身とする。

（2）事務局長・事務局員，監査，学会誌編集委員会委員長及び委員，学会賞選考委員会委員長及び委員の任期は２年とし，引き続き２期までの再任を妨げない。任期は当該年度の４月１日から起算する。

（3）大会実行委員会委員長及び委員の任期は，役員会で承認された日から当該大会に関係する業務の終了時までとする。

（4）顧問を除く役職員の定年を70歳とする。任期の途中で定年に達したときは当該年度の終了まで，その任にあたる。

（役職員の任務）

第14条　役職員の任務は以下の通りとする。

（1）顧問は役員会の求めに応じて学会運営への助言を行う。

（2）事務局長は事務局を主宰し，本会の事務を執行・監督する。

（3）事務局員は事務局長の指示の下，必要な業務を執行する。

（4）監査は本会の会計及び運営が適切になされているか精査し，その結果を総会で報告する。

(5) 学会誌編集委員会委員長は学会誌編集委員会を主宰し，学会誌の刊行にかかる業務を執行・監督する。
(6) 学会誌編集委員は委員長の指示の下，必要な業務を執行する。
(7) 学会賞選考委員会委員長は学会賞選考委員会を主宰し，学会賞・奨励賞の選考にかかる業務を執行・監督する。
(8) 学会賞選考委員は委員長の指示の下，必要な業務を執行する。
(9) 大会実行委員会委員長は大会実行委員会を主宰し，大会の企画・準備・実施にかかる業務を執行・監督する。
(10) 大会実行委員は委員長の指示の下，必要な業務を執行する。

（役職員の選出）
第15条　役職員は会長が推薦し,役員会で承認する。役職員と役員の兼務を妨げない。

（会議）
第16条　本会は以下の会議を開催する。
(1) 総会は会長の招集により，原則として年１回以上開催し，会則の改定，予算・決算その他重要事項を審議する。なお，電子メールやその他の手段を用いた総会の開催も可能とする。総会での議決は出席者の過半数による。
(2) 役員会は会長の招集により，原則として年２回以上開催し，本会の運営にかかる諸問題を審議し，決定する。なお，電子メールやその他の手段を用いた役員会の開催も可能とする。役員会での議決は出席者の過半数による。
(3) 事務局会議は事務局長の判断の下，不定期に開催する。
(4) 学会誌編集委員会，学会賞選考委員会，大会実行委員会は各委員長の判断の下，不定期に開催する。

付則
(1) 本会則は2020年４月１日から施行する。
(2) 本会則は2021年４月１日から改正施行する。
(3) 本会則は2022年４月１日から改正施行する。

（備考）
(1) 本会は1993年４月１日に「英語コーパス研究会」として発足し，1997年４月１日に「英語コーパス学会」に改組されて現在に至る。
(2) 本会の事務局を岐阜市立女子短期大学英語英文学科小島ますみ研究室（〒501-0192 岐阜市一日市場北町７番１号）に置く。

英語コーパス研究（第30号）
【2023年5月31日発行】

編集・発行　2023 英語コーパス学会
〒501-0192 岐阜市一日市場北町 7 番 1 号
岐阜市立女子短期大学英語英文学科　小島ますみ研究室
E-mail: jaecs.hq@gmail.com
Twitter: @JAECS2012
Website: https://jaecs.com/
郵便振替口座：009300-3-195373（英語コーパス学会）
印　刷　所　株式会社ソウブン・ドットコム
〒116-0011 東京都荒川区西尾久 7 丁目 12 番 16 号

ISSN 1340-301 X

English Corpus Studies: Vol.30
2023

Foreword for the Commemorative Issue

Research Articles

Research Notes

Conference Program & Abstract

JAPAN ASSOCIATION FOR ENGLISH CORPUS STUDIES