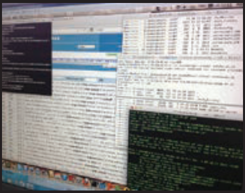
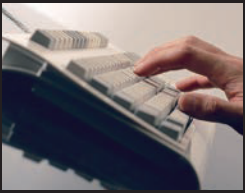
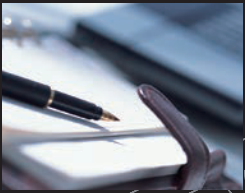
JAECS
Japan Association for English Corpus Studies

2024
英語コーパス学会

ISSN 1340-301 X

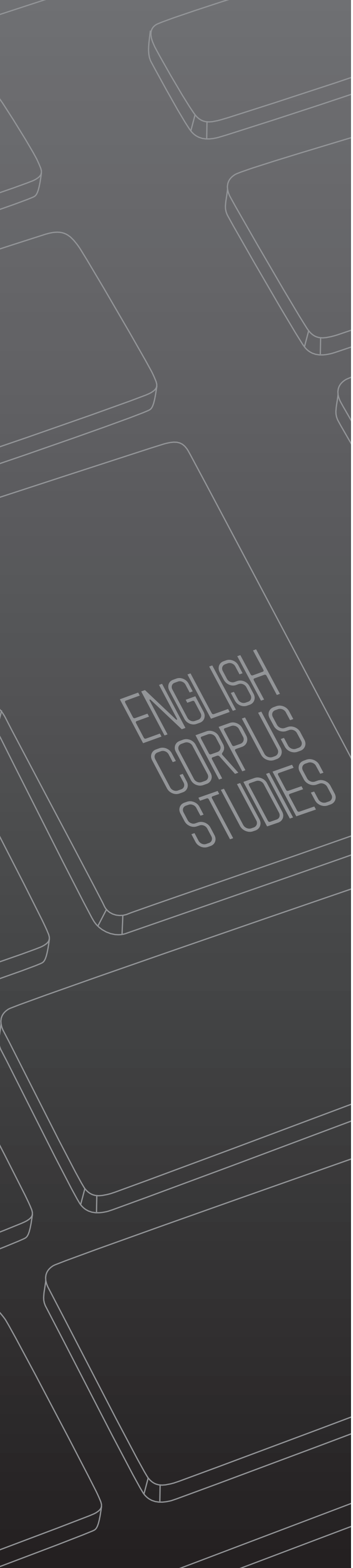






英語コーパス研究
第31号

English Corpus Studies 31



2024年度
英語コーパス学会役員

会　　長：田畑　智司

副 会 長：石井　康毅（研究促進担当）
　　　　　水本　　篤（学会誌担当）
　　　　　小島ますみ（総務担当）

事務局長：小島ますみ

事務局員：後藤　一章，八野　幸子（会計担当）
　　　　　木山　直毅，森下　裕三（広報担当）
　　　　　佐竹　由帆（総務担当）

理　　事：石井　康毅，今林　　修，内田　　諭，岡田　　毅，
　　　　　金澤　俊吾，加野まきみ，小島ますみ，小林雄一郎，
　　　　　杉森　直樹，住吉　　誠，田畑　智司，塚本　　聡，
　　　　　投野由紀夫，水本　　篤，山﨑　　聡（15名）

幹　　事：阿部真理子，五百蔵高浩，石井　達也，石川　有香，
　　　　　和泉　絵美，宇佐美裕子，木山　直毅，佐竹　由帆，
　　　　　長　加奈子，西垣知佳子，仁科　恭徳，八野　幸子，
　　　　　藤原　康弘，森下　裕三，渡辺　拓人（15名）

監　　査：伊藤　亮太

顧　　問：赤野　一郎，中村　純作，堀　　正広

■ 学会誌編集委員会
＜委 員 長＞水本　　篤
＜委　　員＞池尾　玲子，内田　　諭，鎌倉　義士，木山　直毅，
　　　　　　福元　広二，水本　　篤，山﨑のぞみ

■ 学会賞選考委員会
＜委 員 長＞今林　　修
＜委　　員＞今林　　修，長谷部陽一郎，山﨑のぞみ

■ 大会実行委員会
＜委 員 長＞佐竹　由帆
＜委　　員＞佐竹　由帆，石井　康毅，小島ますみ，近藤　悠介，
　　　　　　三浦　愛香，Christopher Cooper

ISSN 1340-301 X

英語コーパス研究

第 31 号

英語コーパス学会

2024

目　次

論文

ソフトウェア紹介

書評

「論文」

A vocabulary study for enhancing learners' experiences: English-language medical research abstracts

Motoko ASANO and Miho FUJIEDA

Abstract

This study investigated the vocabulary features of medical abstracts from the perspective of enhancing learners' vocabulary experiences in the disciplinary field. Our inquiry focused on the prevalence of the General Service List (GSL), the Academic Word List (AWL), the New General Service List (NGSL), and the New JACET List of 8000 Basic Words (hereafter referred to as "New JACET 8000"), as well as the most frequent lexical bundles within these texts. In a corpus of 456,641 tokens with 13,693 types, the repeated use of words and set phrases was found across multiple abstracts despite a high average type/token ratio in the individual texts. The coverages of the GSL and AWL, the NGSL, and the New JACET 8000 were about 80%. The ten highest frequency words accounted for 26% of the total word count with all ten words covered by the GSL; however, most of them were used in context-dependent sequences. The most frequently occurring lexical bundles were highly technical although individual words in the bundles were accessible. These findings may suggest the need for and provide insights into various strategies for raising learners' awareness of the specialized lexical landscape of texts in the disciplinary field.

1. Introduction

1.1 Vocabulary and academic comprehension

The foundation of academic understanding often lies in vocabulary mastery (Coxhead, 2016a; Nation & Macalister, 2021). In academic texts, students frequently encounter barriers due to unfamiliar words or uncertainty regarding their appropriate usage (Charles & Pecorari, 2016; Coxhead, 2017). This challenge is even more pronounced in specialized fields such as medicine (Dang, 2020; Tang & Liu, 2019).

Chung and Nation (2004) revealed that technical vocabulary constituted approximately 30% of an anatomy course text. In addition, students and professionals often have difficulty understanding and retaining the complex terms introduced in medical education and practice (Guest, 2013; Simpson, 2022). Another study suggests that advanced vocabulary learning is necessary for the accurate and appropriate use of terms, which is critical for patient care and interdisciplinary communication (Willey et al., 2019).

Medicine is a field with a high number of students in Japan. There are 81 medical schools or departments (MEXT, 2023a), with about 56,000 students (MEXT, 2023b), indicating one out of 67 could be a medical student based on the total university enrollment of about 630,000 (MEXT, 2023b). This growth was led by the government's 1973 policy addressing physician distribution across the country. Guidelines were issued to help these learners achieve "a level of proficiency" in English to meet the global standards (Hitosugi et al., 2016, p. 88). In these guidelines, "the minimum requirements" aim vocabulary levels to "be able to search for information consisting of English terms and expressions necessary for research in medicine and health care" and their reading skills to "read and understand the abstracts of target English-language research articles" (Japan Society for Medical English Education Guidelines Committee, 2015, pp. 4–5).

The need for teaching basic medical English vocabulary was emphasized by Tamamaki and Fujieda (1998), who revealed the correlation between students' familiarity with essential medical terms and their exposure to medical texts. Shimizu (2019) identified that Japanese medical students in whom the "average TOEIC score was 495 (Range: 270–650)" (p. 83) had difficulty in "understanding the main result of an abstract" (p. 85). These studies underline the complexity of contextual meanings of "lexical items such as *clinical* (compare *clinical trials* with *a clinical decision*)" (Coxhead, 2016b, p. 179). However, our understanding of the vocabulary of medical abstracts from the perspective of enhancing learners' vocabulary experiences is limited. This gap highlights the need for an analysis of the level of vocabulary used in medical research abstracts using word lists created for teaching.

1.2 Vocabulary lists

Teaching the most common words in particular contexts facilitates vocabulary learning (Nation, 2001). Various vocabulary lists have been created to date. Early efforts, such as Thorndike's word book (1921) and an update by Thorndike and Lorge (1944) were followed by lists such as the General Service List (GSL), a contribution "over three decades of work by an international group of leading researchers" (Gilner, 2011, p. 70). The GSL has been "developed from a corpus of 5 million words with the needs of ESL/EFL learners in mind" (Coxhead, 2000, p. 213), with two groups of "998 and 988 word families" (Quero & Coxhead, 2018, p. 54; hereafter, we refer to the two groups as the first one thousand and the first two thousand word families).

The GSL has been used frequently for the development of wordlists. Coxhead prepared the Academic Word List (AWL) of 570 word families (Coxhead, 2000), selected from "a 3,500,000 token corpus of academic English" spanning the Arts, Science, Law, and Commerce (Nation, 2001, p. 188) based on criteria such as the absence in the GSL and the presence in her corpus at least 100 times. The concept of "a word family" is defined as "a headword, its inflected forms, and its closely related derived forms" (Nation, 2001, p. 8). The rationale behind such a family-based organization is articulated by Coxhead (2000, p. 218), who, referencing Bauer and Nation (1993), posits that understanding "regularly inflected or derived members of a family does not require much more effort by learners if they know the base word and if they have control of basic word-building processes." Coxhead and Hirsh (2007) created "*the pilot science corpus*" (p. 70) of 1,761,380 tokens from textbooks and reading materials in the fourteen areas such as "biology" and "sport and health sciences." In their corpus, the coverage of "the first and second thousand of GSL" (Coxhead & Hirsh, 2007, p. 73) was about 70%. Fraser (2007) prepared the Pharmacology Word List (PWL) from a corpus of 51 international pharmacology journal article texts. In his corpus, the coverage of the GSL and AWL was 70.44% (Fraser, 2007). Wang et al. (2008) created "a Medical Academic Word List" (p. 445) using a corpus of over 1.09 million words from research articles of "almost all the fields of medical science" (p. 445) for designing curriculum in medical English education (Wang et al., 2008). Fraser (2007) also compiled a 58,413-token corpus of a pharmacology textbook, in which the coverage of the GSL and AWL was 68.77% (Fraser, 2007). Quero and Coxhead (2018) prepared two corpora including "a medical corpus" of two medical textbooks that cover "a comprehensive range of medical topics" (p. 58) and "a second medical corpus" of medical textbooks covering "a wide range of medical topics" (p. 59). In their corpora, the GSL coverage was about 60% with the AWL coverage being around 8%. Chen and

Ge (2007) created a corpus of “50 medical research articles with 190,425 running words” (Chen & Ge, 2007, p. 506) and examined the top 20 academic word items from “abstract,” “introduction,” “materials and methods,” “results,” and “discussion” sections. In their entire corpus, the coverage of the AWL was “10.073%” (Chen & Ge, 2007, p. 508), with the AWL words occurring evenly in the sections. The AWL was also used to develop a word list for medical professionals and students based on a corpus of 99 research articles (Tang & Liu, 2019).

The combination of the GSL and AWL has been considered “of relevance” (Gilner, 2011, p. 74); however, “the use of the 50-year-old GSL” (Green & Lambert, 2018, p. 107) has been criticized. “A word family approach” (Culligan, 2019, p. 37) has been regarded as “problematic” (Gardner & Davies, 2014, p. 307). The New General Service List (NGSL, Browne, 2013) was “conceived as a modern update of the General Service List (West, 1953)” (Mizumoto et al., 2021, p. 31). The NGSL is considered “optimal” for Japanese learners (Mizumoto et al., 2021, p. 32; Nakata, 2022, p. 23). According to Culligan’s study (2019), Japanese undergraduates’ perspectives on the GSL and NGSL suggest that the NGSL may offer a slightly easier learning curve and would be more suitable for learners.

In contrast, the New JACET 8000 (JACET Special Committee for Revision of the JACET Wordlist, 2016) is “an updated version of the JACET 8000 word list (JACET Committee for Revision of the JACET Wordlist, 2003). This wordlist was compiled by the Japan Association of College English Teachers (JACET)” (Mizumoto et al., 2021, p. 31). The updated version is regarded as “a list of the 8000 basic words which should be acquired by learners of English” (Terauchi, 2016, p. 13) and is, therefore, primarily for a Japanese audience; it includes lemmatized words with part-of-speech information.

1.3 Lexical bundles

Corpus studies have increasingly identified recurring “communicative events” (Swales, 1990, p. 9), focusing on prefabricated word sequences within their contexts. Lexical bundles, defined as “sequences of word forms” (Biber et al., 1999, p. 990), occur more commonly across language events than would be expected by chance. These “multi-word sequences” (Biber et al., 2004, p. 373; Mizumoto, 2015, p. 30) can be identified as “n-grams” (Mizumoto, 2015, p. 31; Stubbs & Barth, 2003, p. 61). It has been shown that 4-word bundles "hold 3-word bundles in their structure" (Cortes, 2004, p. 401; Hyland, 2008, p. 6) and are "far more common than 5-word strings" (Hyland, 2008, p. 8). Stubbs and Barth (2003) argue that "n-grams," referred to as "chains of word-forms" (p. 61), "are not necessarily linguistic units" (p. 62); however, frequently occurring 4-grams are shown to characterize text types in such a way that research articles have frequently occurring 4-grams such as "*on the other hand* and *at the beginning of*"(Cortes, 2013, p. 34). Examining n-grams is considered to "complement measures (such as type-token ratio) which can characterize text-types" (Stubbs & Barth, 2003, p. 79). The need for "teaching lexical bundles" in "English for Academic/Specific Purposes (EAP/ESP)" (Mizumoto, 2015, p. 33) settings has been underscored.

The approach has been applied to examine medical research articles. Jalali et al. (2015) identified that "in the present study" (p. 57) was the most frequent 4-gram in their 2.4-million-word research article texts from 33 "medical subject areas" (p. 54). Abdollahpour and Gholami (2018) identified "all four-word lexical bundles" (p. 90) within their corpus of "the abstract sections" (p. 82) from various medical journal articles, with subsequent "categorization into two major groups of general and technical" (p. 90). In their corpus, "this study was to" (p. 105) was "the most frequent general lexical bundle" (Abdollahpour & Gholami, 2018).

1.4 Medical research writing

Medical articles, especially in health research, must adhere to specific study design requirements, as reported by Millar et al. (2019). An international initiative that seeks to "ensure quality in the reporting" (Millar et al., 2019, p. 141) now provides "616 reporting guidelines" (EQUATOR Network, 2024) that stipulates "the required sections and information" (Millar et al., 2019, p. 150) necessary for each publication. Non-compliance with these guidelines "may result in a manuscript being deemed of inferior quality" (Millar, et al., 2019, p. 141) by the International Committee of Medical Journal Editors (ICMJE), working "to improve the quality of medical science and reporting" (ICMJE, 2024) and issuing "similar guidelines for medical research articles in general (the updated recommendations may be found at www.icmje.org)" (Millar et al., 2019, p. 141). Their recommendations are "widely accepted by biomedical journals" (Luo & Hyland, 2019, p. 39), and "play a central role" (Millar et al., 2012, p. 393) in research writing.

1.5 Aim of this study and research questions

This study aims to examine the vocabulary in our corpus of English-language abstracts of medical research articles. By examining our corpus texts with authentic vocabulary lists, we try to understand the vocabulary level of medical research abstracts. The present study poses the research questions (RQs) “What is the prevalence of words from the GSL, AWL, NGSL, and the New JACET 8000 within the corpus texts?” and “What are the most frequent lexical bundles in the corpus texts?”

2. The corpus and previous findings

2.1 The corpus

Our corpus comprises 1,481 abstracts of research articles from an international journal, *the New England Journal of Medicine* (*NEJM*), published in 2010 and 2015 through 2020. The gap in the years was due to a shortfall of human resources, which prevented the incorporation of abstracts from missing years. We used this journal’s abstracts because the publications meet the criteria of “representativity, reputation, and accessibility” (Nwogu, 1997, p. 121), offer canonical insights for Japanese students in their training to “evaluate medical literatures” (Ogawa, 2014, p. 41) and also in learning how a vocabulary item “is actually used in writing for medical professionals” (Jego, 2012, p. 51). The journal offers official translations in Japanese (The New England Journal of Medicine, 2024a). The texts from the website were extracted, segmented into sentences, organized in spreadsheet columns, and saved as individual abstract files.

2.2 Previous findings from the corpus

Our previous study examined the 2018 part of this corpus. In the study, modal verbs appear mainly in the Introduction of the abstracts, followed by the Conclusion, Methods, and Results sections. The majority of the top 20 collocates for the “collocational framework” (Renouf & Sinclair, 1991, p. 128) “the . . . of” (Marco, 2000, p. 63) match those found in Marco’s list such as “the risk of,” “the effect of,” and “the presence of.” Punctuation marks like commas, semicolons, or colons have overlapping functions (Asano et al., 2021).

3. Methods

3.1 Wordlists

Our analysis focused on the vocabulary profiles of the texts in the entire corpus, concentrating on the coverage by general and academic word lists. For our analysis with the GSL (West, 1953) and the AWL (Coxhead, 2000), we relied on versions of these lists that were "created by Paul Nation and cleaned by Laurence Anthony" (Anthony, n.d.). We used the NGSL (Version 1.2; Browne, 2013) by accessing the website for the list. When using the New JACET 8000 in our analysis, we employed the concept of "flemma," defined as "a base form as a headword and its inflected forms as one word," a counting approach that "combines inflections of lemma groups but does not distinguish the POS" (Mizumoto et al., 2021, p. 33). We examined our corpus texts in their original forms for analyses involving other wordlists.

3.2 Tools

The texts were examined with CasualConc (Version 3.0.6; Imao, 2023), AntConc (Version 4.2.4; Anthony, 2023), and AntWordProfiler (Version 1.5.1; Anthony, 2021). The quantitation of vocabulary coverage deployed the stopword function of CasualConc. The target items were "removed in the pre-processing step" (Sarica & Luo, 2021, p. 1). The word frequency after removal was subtracted from the total word count of the texts to determine the word count. The results were processed using Microsoft Excel (version 2308) and Google Colaboratory's Python environment. AntWordProfiler was used to determine the results and quantify types in each text. The data was lemmatized using spaCy (version 3.6.1) in the Python environment where necessary. AntConc was also used to obtain "n-grams (or "lexical bundles")" to identify multi-word expressions that characterize the specific texts (Nesi, 2013, p. 418).

4. Results

4.1 Word profiles of the corpus texts

The corpus contained 456,641 words of 13,693 types (Table 1). The average word count per abstract exceeded 300 after 2016, with an increasing variation in length over time.

Table 1. Word count of the corpus texts per year

Year	2010	2015	2016	2017	2018	2019	2020	Total
Number of abstracts	208	223	208	208	208	209	217	1,481
Type	4,921	5,046	4,992	4,969	4,876	4,928	5,004	13,693
Total word count (Token)	58,747	65,377	65,290	65,726	65,684	67,736	68,081	456,641
Mean word count	282	293	314	316	316	324	314	308
Standard deviation	30	32	41	40	38	39	36	39

According to the journal guidelines (The New England Journal of Medicine, 2024b), a research article should not exceed 2,700 words, including the abstract, a maximum of five tables and figures, and up to 40 references. The guidelines do not set a word count limit for the abstracts, but an upward trend in the average word count was seen over the study period, accompanied by greater variation among the abstracts.

The type/token ratio (TTR) of individual abstract texts averaged around 45.0 (Figure 1). This ratio was high, considering that "general prose and essays in British and American English" texts in the Freiburg LOB (FLOB) and Freiburg-Brown (Frown) corpora have a TTR of "8.14" (Fujiwara, 2003, p. 93). These findings may be attributable to "the use of many different lexical items in a text" (Biber, 1988, p. 104). However, the entire corpus showed 13,693 types and 456,641 tokens (Table 1), with the top 100 words accounting for 53.1% of the total frequency (Table 2), suggesting

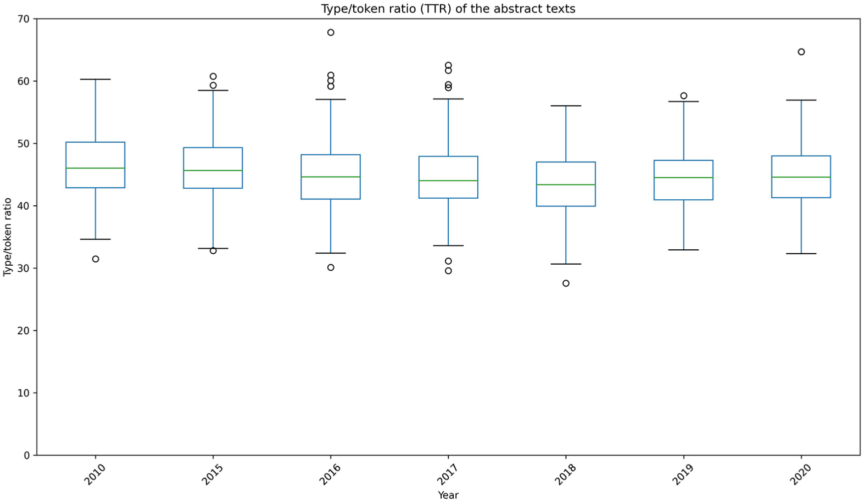


Figure 1. Distribution of type/token ratios (TTRs) for the individual abstract texts.

Table 2. Top 100 words in the corpus

	Word	Raw freq.	Std (%)		Word	Raw freq.	Std (%)		Word	Raw freq.	Std (%)		Word	Raw freq.	Std (%)		Word	Raw freq.	Std (%)
1	the	24,564	5.4	21	placebo	2,306	0.5	41	on	1,364	0.3	61	higher	960	0.2	81	study	793	0.2
2	of	20,186	4.4	22	from	2,175	0.5	42	therapy	1,351	0.3	62	no	947	0.2	82	median	791	0.2
3	in	16,366	3.6	23	than	2,099	0.5	43	not	1,344	0.3	63	between	927	0.2	83	age	784	0.2
4	and	13,222	2.9	24	among	2,067	0.5	44	trial	1,341	0.3	64	compared	915	0.2	84	during	773	0.2
5	to	10,906	2.4	25	as	1,897	0.4	45	funded	1,327	0.3	65	those	902	0.2	84	significantly	773	0.2
6	with	9,084	2.0	26	per	1,860	0.4	46	death	1,300	0.3	66	this	891	0.2	86	control	768	0.2
7	a	8,249	1.8	27	that	1,766	0.4	47	assigned	1,288	0.3	67	days	886	0.2	87	difference	745	0.2
8	patients	6,963	1.5	28	ratio	1,627	0.4	48	dose	1,226	0.3	67	lower	886	0.2	88	clinical	723	0.2
9	was	5,992	1.3	29	an	1,609	0.4	49	survival	1,214	0.3	69	participants	885	0.2	89	cancer	721	0.2
10	group	5,453	1.2	30	primary	1,586	0.3	50	end	1,137	0.2	70	occurred	882	0.2	90	groups	720	0.2
11	were	4,903	1.1	31	risk	1,579	0.3	51	is	1,133	0.2	71	outcome	873	0.2	91	health	714	0.2
12	or	4,796	1.1	32	treatment	1,548	0.3	52	interval	1,118	0.2	72	total	872	0.2	92	rates	704	0.2
13	for	4,608	1.0	33	after	1,473	0.3	53	received	1,068	0.2	73	receive	866	0.2	93	free	694	0.2
14	at	3,205	0.7	34	mg	1,464	0.3	54	associated	1,067	0.2	74	more	853	0.2	94	up	692	0.2
15	by	2,696	0.6	35	number	1,442	0.3	55	adverse	1,066	0.2	75	randomly	848	0.2	95	overall	691	0.2
16	p	2,683	0.6	36	events	1,427	0.3	56	confidence	1,047	0.2	76	weeks	844	0.2	96	outcomes	681	0.1
17	ci	2,669	0.6	37	disease	1,423	0.3	57	gov	1,007	0.2	77	two	826	0.2	97	all	680	0.1
18	who	2,488	0.5	38	years	1,399	0.3	58	clinicaltrials	1,005	0.2	78	point	822	0.2	98	percentage	674	0.1
19	we	2,450	0.5	39	rate	1,398	0.3	59	hazard	1,002	0.2	79	points	799	0.2	99	care	673	0.1
20	had	2,365	0.5	40	months	1,389	0.3	60	vs	970	0.2	79	response	799	0.2	100	year	667	0.1

Raw freq: Raw frequency; Std (%): Frequency per one hundred words

that the texts contain many repetitive words and phrases.

The ten most frequent words accounted for 26% of the tokens. They were covered by the GSL but often used in context-dependent sequences. For example, the two most frequently used words, “the” and “of,” appeared in 455, 500, and 183 instances of four-grams “the . . . of the,” “in the . . . of,” and “the . . . of a,” respectively. There were 341, 108, and 61 instances of “the risk of,” “the effect of,” and “the presence of,” respectively; these are the top three “collocates” of the “the . . . of” framework in Marco’s study (2000, p. 68). These findings underscore the significant presence of this framework in our corpus. Words such as “per,” “ratio,” and “rate” frequently recurred, expressing “measure [and] quantification” (Marco, 2000, p. 69).

4.2 Lexical coverage of the wordlists

In the corpus text, the coverage of the first one thousand word families of the GSL was 59.4% on average, ranging from 39.1% to 80.4%. The coverage of the first two thousand word families was 64.7% and ranged from 43.9% to 84.9%. The coverage of the GSL and AWL was 75.2%, ranging from 54.7% to 95.0%. The NGSL showed a mean coverage of 75.1%, having a significantly greater coverage compared to the first and second two thousand word families of the GSL ($t = 51.162$, $df = 2{,}960$, $p < .01$). The cumulative coverage of the New JACET 8000 averaged 83.2%, ranging from 63.7% to 96.8% (Figure 2). The last 1,000 words in the word lists include specific terms such as “aspirin,” “variant,” “pneumonia,” “artery,” “coronary,” “infusion,” “renal,” “cardiovascular,” and “tuberculosis.” However, no marked difference was seen between the cumulative coverage of the 7,000 words and that of 8,000 words (Figure 2), suggesting that the last 1,000 words did not significantly contribute to the overall coverage.

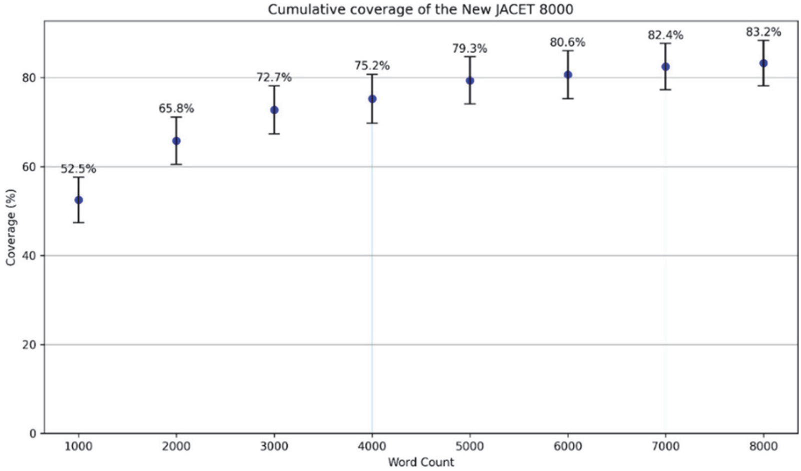


Figure 2. The mean and standard deviation of the cumulative coverage of the New JACET 8000.

4.3 Most frequent lexical bundles in the corpus

The ten most frequent 3-grams, 4-grams and 5-grams (Table 3) had a standard frequency of over 300 instances per million words. All ten most frequent 4-grams had a range of more than 10%. The instances of the bundles were noticeably frequent, taking into consideration the cut-off frequency of “20 per million words” (Cortes, 2004, p. 400; Hyland, 2008, p. 9; Hyland & Jiang, 2018, p. 385) used for quantifying bundles.

The most frequently occurring 3-gram “ci to p” was identified as “part of a 5-word string” (Hyland & Jiang, 2018, p. 386) “confidence interval ci to p” (Figure 3). The 4-gram “interval ci to p” “hold” (Cortes, 2004, p. 401) the three-word bundles “ci to p,” “confidence interval ci,” and “interval ci to” as shown in an example corpus text (Glauser et al., 2010, p. 790): “After 16 weeks of therapy, the freedom-from-failure rates for ethosuximide and valproic acid were similar (53% and 58%, respectively; odds ratio with valproic acid vs. ethosuximide, 1.26; 95% confidence interval [CI], 0.80 to 1.98; P = 0.35) and were higher than the rate for lamotrigine (29%; odds ratio with ethosuximide vs. lamotrigine, 2.66; 95% CI, 1.65 to 4.28; odds ratio with valproic acid vs. lamotrigine, 3.34; 95% CI, 2.06 to 5.42; P < 0.001 for both comparisons)." The frequent use of "confidence interval" and "p" values suggests their role in describing study findings.

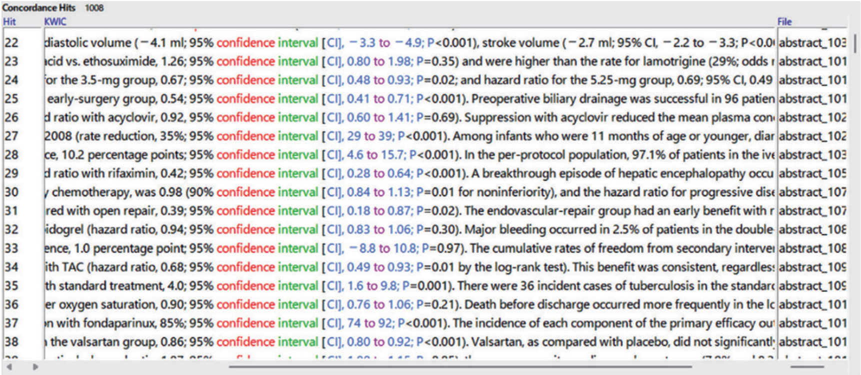
The most frequent 4-gram "clinical trials gov nct," which occurred in 62.8% of the texts, refers to a unique identifier assigned to clinical trials registered on a database of "clinical research studies" conducted around the world (ClinicalTrials.gov, 2023). This identifier, known as the "National Clinical Trial (NCT) number," is part of a global registry network that facilitates access to trial information and shows that the study satisfies the recommendation "as a condition of consideration for publication" (ICMJE, 2024, p. 13), highlighting the significance of referencing the regulatory aspects of research in this corpus text.

Of the 379 instances of the 4-gram "primary end point was," 370 (97.6%) occurred in the second most frequent 5-gram "the primary end point was." All these words are in the NGSL; "primary" is in the AWL and ranked at the 1098th in the New JACET 8000; "end" and "point" are both in the first thousand words of the GSL and rank in the 180s in the New JACET 8000. How-

Table 3. The most frequently occurred lexical bundles

Rank	Three-gram	Raw	Std.	Range	Four-gram	Raw	Std.	Range	Five-gram	Raw	Std.	Range
1	ci to p	1,008	2,207	33.3	cinicaltrials gov number nct	932	2,041	62.8	confidence interval ci to p	398	872	26.9
2	patients in the	965	2,113	28.6	confidence interval ci to	767	1,680	51.8	the primary end point was	370	810	23.8
3	gov number nct	935	2,048	63.0	in the placebo group	636	1,393	18.8	at the dose of mg	224	491	11.4
4	clinicaltrials gov number	932	2,041	62.8	the primary end point	482	1,056	26.7	ratio confidence interval ci to	218	477	15.1
5	confidence interval ci	773	1,693	52.2	interval ci to p	398	872	26.9	hazard ratio ci to p	176	385	10.3
6	interval ci to	767	1,680	51.8	hazard ratio ci to	382	837	15.3	of the patients in the	152	333	7.8
7	the placebo group	753	1,649	20.0	primary end point was	379	830	24.3	in a ratio to receive	149	326	9.9
8	as compared with	720	1,577	33.0	group and in the	326	714	16.6	hazard ratio confidence interval ci	149	326	10.1
9	funded by the	718	1,572	48.2	the primary outcome was	317	694	20.4	in of patients in the	149	326	6.1
10	in the placebo	668	1,463	19.6	a total patients	316	692	20.3	primary end point was the	148	324	9.9

Raw: Raw frequency; Std.: Frequency per one million words; Range: Percentage



Concordance Hits 1008

Hit	KWIC	File
22	diastolic volume (−4.1 ml; 95% confidence interval [CI], −3.3 to −4.9; P<0.001), stroke volume (−2.7 ml; 95% CI, −2.2 to −3.3; P<0.0	abstract_103
23	acid vs. ethosuximide, 1.26; 95% confidence interval [CI], 0.80 to 1.98; P=0.35) and were higher than the rate for lamotrigine (29%; odds r	abstract_101
24	for the 3.5-mg group, 0.67; 95% confidence interval [CI], 0.48 to 0.93; P=0.02; and hazard ratio for the 5.25-mg group, 0.69; 95% CI, 0.49	abstract_101
25	early-surgery group, 0.54; 95% confidence interval [CI], 0.41 to 0.71; P<0.001). Preoperative biliary drainage was successful in 96 patien	abstract_101
26	d ratio with acyclovir, 0.92, 95% confidence interval [CI], 0.60 to 1.41; P=0.69). Suppression with acyclovir reduced the mean plasma con	abstract_102
27	2008 (rate reduction, 35%; 95% confidence interval [CI], 29 to 39; P<0.001). Among infants who were 11 months of age or younger, diar	abstract_102
28	ce, 10.2 percentage points; 95% confidence interval [CI], 4.6 to 15.7; P<0.001). In the per-protocol population, 97.1% of patients in the ive	abstract_103
29	d ratio with rifaximin, 0.42; 95% confidence interval [CI], 0.28 to 0.64; P<0.001). A breakthrough episode of hepatic encephalopathy occu	abstract_105
30	y chemotherapy, was 0.98 (90% confidence interval [CI], 0.84 to 1.13; P=0.01 for noninferiority), and the hazard ratio for progressive dise	abstract_107
31	red with open repair, 0.39; 95% confidence interval [CI], 0.18 to 0.87; P=0.02). The endovascular-repair group had an early benefit with r	abstract_107
32	idogrel (hazard ratio, 0.94; 95% confidence interval [CI], 0.83 to 1.06; P=0.30). Major bleeding occurred in 2.5% of patients in the double-	abstract_108
33	ence, 1.0 percentage point; 95% confidence interval [CI], −8.8 to 10.8; P=0.97). The cumulative rates of freedom from secondary interve	abstract_108
34	ith TAC (hazard ratio, 0.68; 95% confidence interval [CI], 0.49 to 0.93; P=0.01 by the log-rank test). This benefit was consistent, regardless	abstract_109
35	th standard treatment, 4.0; 95% confidence interval [CI], 1.6 to 9.8; P=0.001). There were 36 incident cases of tuberculosis in the standar	abstract_109
36	er oxygen saturation, 0.90; 95% confidence interval [CI], 0.76 to 1.06; P=0.21). Death before discharge occurred more frequently in the lo	abstract_101
37	n with fondaparinux, 85%; 95% confidence interval [CI], 74 to 92; P<0.001). The incidence of each component of the primary efficacy ou	abstract_101
38	the valsartan group, 0.86; 95% confidence interval [CI], 0.80 to 0.92; P<0.001). Valsartan, as compared with placebo, did not significantl	abstract_101

Figure 3. The concordance lines showing the 3-gram “ci to p”

ever, the combined sequence “primary end point” refers to “the study’s objective” (ICMJE, 2024, p. 17), indicating that the term is used to denote the “research outcome” (Nwogu, 1997, p. 132). Although each word may be accessible for learners, the sequence “primary end point” is highly technical. This gap between the general accessibility of individual words and the specialized use of the word sequence highlights the complexity of the texts.

5. Discussion

Our analysis revealed the vocabulary diversity and the repeated use of high frequency words, showing bundles to be highly technical. Individual texts had many word types, as shown by a mean TTR of about 45.0 (Figure 1). In contrast, the entire corpus contained 13,693 types and 456,641 tokens (Table 1), with the top 100 words accounting for 53.1% of the total frequency (Table 2). TTRs are affected by “the size of the corpus” (McEnery & Hardie, 2012, p. 50), but our corpus data showed the repetitive use of words and set phrases across multiple abstracts. This was consistent with the finding that the range, or “document frequency” (Tabata, 2012, p. 3) divided by the total number of texts, scored 50% or greater in several and exceeded 10% in many word sequences (Table 3), indicating conventional terminology usage across the corpus. The bundles were quite technical although each word was commonly found in the wordlists. These findings highlight the needs for furnishing students with lexical tools to navigate these texts effectively.

The first research question (RQ) explored the prevalence of words from specific lists within the corpus, finding mean coverages of 75.2%, 75.1%, and 83.2% for the GSL and AWL, NGSL, and New JACET 8000, respectively, highlighting the vocabulary complexity for learners. Nation and Macalister (2021) note understanding necessitates familiarity with at least "98 per cent" (p. 12) of text vocabulary.

Addressing the second RQ on frequent lexical bundles, we found the leading 3-gram "ci to p" within the 5-gram "confidence interval ci to p" appeared 872 times per million words across 26.9% of texts (Table 3), illustrating its use in stating study findings. Despite individual words in the bundle such as "primary end point was" being accessible, the sequence should be notably technical, underscoring textual complexity.

Our study needs to consider our target learners' vocabulary levels. Hamada et al. (2021) recorded an average vocabulary size of 4,575 words among over 1,000 "students from 16 Japanese universities (29 faculties)" (p. 29) using the New JACET 8000 list. Beglar (2010), using the vocabulary size test by Nation (2006), found vocabulary sizes ranging between 4,700 and 5,700 for Japanese university students of varying English proficiency levels. McLean et al. (2014) reported an average vocabulary size of "3,939 word families" (p. 35) with the "Vocabulary Size Test" (p. 34) by Nation and Beglar (2007). Although we must interpret these results individually and cautiously, the findings imply that the learners may not fully grasp all vocabulary present in the word lists examined in our study.

The major approach, among vocabulary learning strategies, is introduced by Nation and Macalister (2021). They advocate for "intensive reading" (p. 42), where teachers can help "learners use context clues to guess the meaning of the word" (p. 43).

To improve medical students' English proficiency, Fraser et al. (2015) developed "word lists" (p. 16) from medical texts, including doctor-patient conversation and "an anatomy textbook *Gray's Anatomy for Students*" (p. 17). Fraser et al. (2015) hypothesized that students' familiarity with the content "would help them greatly when they encountered difficult words or sentences" (p. 18). They integrated corpus studies with their classroom activities, facilitating students' association with the subject-matter contexts (Fraser et al., 2015). The creation of word lists, informed by interviews with medical professors and "feedback from doctors" (Fraser et al., 2015, p. 18), received "positive feedback" (p. 19) from students.

An alternative learning strategy involves leveraging learners' first language (L1) to establish the "meaning-form link" (Schmitt, 2008, p. 353) despite discouragement from the Ministry of Education, Culture, Sports, Science and Technology's policy (MEXT, 2014). A survey shows learners' preference for "the idea of F[oreign] L[anguage] learning as bilingual education" (Turnbull, 2018, p. 119), suggesting that "translanguaging" (García, 2009, p. 45; Turnbull, 2018, p. 101), could be a mainstream option. García (2009, p. 45) posits that the term "translanguaging" as bilinguals' use of their languages to become aware of their multilingual worlds, a concept originated from Cen Williams (Baker, 2001).

New studies highlight the benefits of translanguaging in higher education. Shoecraft et al. (2024) reported on their action research in a first-year anatomy course at an Australian university, where more than 30% of students were learning English as an additional language (EAL). The translation of mini-lecture video transcripts into the students' first languages proved beneficial as scaffolding. Scanning in the first language before reading or viewing in English saved time in understanding the content and helped to build students' confidence. Zheng and Drybrough (2023) investigated the translanguaging practices of five Chinese postgraduate students during the outlining, note-taking, and drafting stages of their master's dissertation writing process at a British university. The study reveals six translanguaging practices, such as "to illustrate the relationship between different pieces of information" (Zheng & Drybrough, 2023, p. 9), supported students' self-regulation and efficiency in controlling the extensive writing process to achieve their writing goals. In a mixed-methods study, Galante (2020) investigated the effects of translanguaging on academic vocabulary development compared to a traditional monolingual approach. The results of vocabulary tests, classroom observations, and learner diaries at the end of the 12-week EAP program revealed that the translanguaging group had a significantly higher academic vocabulary than the monolingual group. Active engagement in cross-linguistic meaning making was observed in the translanguaging group.

In classrooms sharing "similar L1-related difficulties" (Flowerdew, 2012, p. 215), "a bilingual corpus" (Aijmer, 2002, p. 1) containing "source texts and their translations" (Baker, 1993, p. 248) aids in exploring concordance lines "as students can see the different contexts in which a word is used" (Flowerdew, 2012, p. 215). Chujo et al. (2006) used an online bilingual concordancer equipped with "Japanese-English parallel corpora" (p. 153) of news articles in Japanese university beginning-level English classes for activities like identifying patterns and tendencies. Using a concordancer helped learners find language patterns "themselves (with guidance from handouts)" (Chujo et al., 2006, p. 169).

Our project developed the Medical English Education Support System (MEE-SUS) featuring bilingual concordancing from our corpus texts and the official Japanese translations (Nakano et al., 2021). The journal's regional site (Nankodo, 2024) offers Japanese translations of abstracts of various "original articles" (The New England Journal of Medicine, 2024b). In one study, about 100 first-year medical students in a private university were asked to become familiar with the system, examine language items the participants picked up from the tool, and write their findings on a worksheet in a required course (Asano et al., 2022a). These students averaged 475.04 in TOEFL ITP score with a standard deviation (SD) of 43.17, suggesting that most students were at or below the CEFR B1 level (Oshimi, 2022). They were subsequently involved in "the peer-worksheet viewing activity" (Asano et al., 2022a, p. 22), and many commented their surprise in learning the contextual meaning of items such as "mean" for "heikin," "case" for "shorei," and "develop" for "shojiru [to occur]." One participant reported, "I was surprised to learn that "subject" has the meaning of "hikensha." I will keep this in mind as I will be using it a lot in the future" (Asano et al., 2022a, p. 24). In another study, fourth-year medical students (average TOEFL ITP score 455.9; SD 45.7) were introduced to "guidelines" (Millar et al., 2019, p. 150), read a model abstract to review how the information was given in a required course at the same university (Asano et al., 2022b). They were tasked to choose an abstract, extract information such as "trial design," and write a summary. Those who used the bilingual display of the tool scored higher than non-users in all task items. Although the "learner-directed corpus projects" have invited arguments such as "whether such an approach is feasible is questionable" (Ballance & Coxhead, 2022, p. 412), these attempts might foster learners' "awareness" and "tolerance" foreseeable in their "real world" community (Cook, 2010, p. 117–118).

This study had some limitations: it analyzed only abstracts from the past seven years. With ongoing data addition, a full-scale of analysis may be optimal in the future. This study used the New JACET 8000, which contains lemmatized words with part-of-speech information. However, the corpus texts were lemmatized in the analysis without considering the part-of-speech information.

The corpus texts exhibited specialized vocabulary and lexical bundles suggestive of disciplinary conventions. Nation and Macalister (2021, p. 137) propose that an English course incorporating digital tools could be "a means of improving and developing information gathering skills in both L1 and L2." Nation (2001) argues the difficulty of technical vocabulary for a disciplinary novice learner guessing from the context as "the reader does not already have a good background in that technical area" and thus "looking the word up in a dictionary does not bring much satisfaction" (p. 204). In a meta-analysis, the use of parallel corpora was found to be effective for learners whose first language is Japanese (Boulton & Cobb, 2017). Although a broader discussion is needed, the findings of this study may suggest the need for and provide insights into different strategies for helping students to "succeed in the learning contexts" (Tribble, 2017, p. 40) and raising learners' awareness of specialized language used in medical abstracts.

Acknowledgments

This study was supported by Grant-in-Aid for Scientific Research by the Japan Society for the Promotion of Science (JSPS), with Grant Numbers: 18K02966 and 23K02800. We would like to express our sincere appreciation to the insightful comments from the reviewers.

References

Abdollahpour, Z., & Gholami, J. (2018). Building blocks of medical abstracts: Frequency, functions and structures of lexical bundles. *The Asian ESP Journal, 14*(1), 82–110.

Aijmer, K. (2002). What can translation corpora tell us about discourse particles? *English Corpus Studies, 9*, 1–15.

Anthony, L. (n.d.). AntWordProfiler homepage. https://www.laurenceanthony.net/software/antwordprofiler

Anthony, L. (2021). AntWordProfiler (Version 1.5.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2023). Help for AntConc (Windows, MacOS, Linux) Build 4.2.4. Available from https://www.laurenceanthony.net/software/antconc

Asano, M., Fujieda, M., & Noguchi, J. (2021). Linguistic and punctuational features of research article abstracts in English-Japanese parallel corpora—Envisaging pedagogical applications. *Proceedings of INTED2021 Conference 8th–9th March 2021*, 3287–3296.

Asano, M., Nakano, M., Miyazaki, Y., & Fujieda, M. (2022a). Introducing a bilingual corpus database system of medical abstracts for exploring academic connotations of words: A case study of first-year medical students. *Journal of Medical English Education, 21*(1), 18–26.

Asano, M., Nakano, M., Miyazaki, Y., Wakasa, T., & Fujieda, M. (2022b). Use of authentic translation in helping students decipher English-language randomised control trial abstracts. *Journal of Medical English Education, 21*(3), 87–93.

Baker, C. (2001). *Foundations of bilingual education and bilingualism* (3rd ed.). Multilingual Matters.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–352). John Benjamins.

Ballance, O., & Coxhead, A. (2022). What can corpora tell us about EAP. In A. O'Keeffe (Ed.), *The Routledge handbook of corpus linguistics* (2nd ed., pp. 405–415). Routledge.

Bauer, L., & Nation, P. (1993). Word Families. *International Journal of Lexicography*, *6*(4), 253–279.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118.

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at . . .: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

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

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

Browne, C. (2013). The New General Service List: Celebrating 60 years of vocabulary learning. *The Language Teacher, 37*(4), 13–16.

Charles, M., & Pecorari, D. (2016). *Introducing English for academic purposes.* Routledge.

Chen, Q., & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes, 26*(4), 502–514.

Chujo, K., Utiyama, M., & Miura, S. (2006). Using a Japanese-English parallel corpus for teaching English vocabulary to beginning-level students. *English Corpus Studies, 13*, 153–172.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System, 32*(2), 251–263.

ClinicalTrials.gov. (Last updated on May 23, 2023). About ClinicalTrials.gov. Retrieved from https://clinicaltrials.gov/about-site/about-ctg

Cook, G. (2010). *Translation in language teaching*. Oxford University Press.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397–423.

Cortes, V. (2013). *The purpose of this study is to:* Connecting lexical bundles and moves in research article introductions. *English for Specific Purposes, 12*, 33–43.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Coxhead, A. (2016a). Academic vocabulary. In M. Charles & D. Pecorari (Eds.), *Introducing English for academic purposes* (pp. 108–121). Routledge.

Coxhead, A. (2016b). Acquiring academic and disciplinary vocabulary. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 177–190). Routledge.

Coxhead, A. (2017). Approaches and perspectives on teaching vocabulary for discipline-specific academic writing. In J. Flowerdew & T. Costley (Eds.), *Discipline-specific writing* (pp. 62–76). Routledge.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée, 2*, 65–78.

Culligan, B. A. (2019). Evaluating corpora with word lists and word difficulty. *Vocabulary learning and instruction, 8*(1), 29–38.

Dang, T. N. Y. (2020). The potential for learning specialized vocabulary of university lectures and seminars through watching discipline-related TV programs. *TESOL Quarterly, 54*(2), 436–459.

EQUATOR Network. (2023). *Reporting guidelines*. Retrieved from http://files.eric.ed.gov/fulltext/ED332551.pdf

Flowerdew, L. (2012). *Corpora and language education.* Palgrave Macmillan.

Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education, 10*, 127–143.

Fraser, S., Davies, W., & Tatsukawa, K. (2015). Creating a corpus-informed EMP course for medical undergraduates. *Journal of the IATEFL ESP SIG*, 16–21.

Fujiwara, Y. (2003). The use of reason-consequence conjuncts in Japanese learners' written English. *English Corpus Studies, 10*, 91–104.

Galante, A. (2020). Translanguaging for vocabulary development: A mixed methods study with international students in a Canadian English for academic purposes program. In Z. Tian, L. Aghai, P. Sayer, & J. L. Schissel (Eds.), *Envisioning TESOL through a Translanguaging lens: Global perspectives* (pp. 293–328). Springer Nature.

García, O. (2009). *Bilingual education in the 21st century*. Wiley-Blackwell.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. *Applied Linguistics, 35*(3), 305–327.

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language, 23*(1), 65–83.

Glauser, T. A., Cnaan, A., Shinnar, S., . . . & Adamson, P. C. for the Childhood Absence Epilepsy Study Group. (2010). Ethosuximide, valproic acid, and lamotrigine in childhood absence epilepsy. *The New England Journal of Medicine, 362*, 790–799.

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. *English for Specific Purposes, 35*, 105–115.

Guest, M. (2013). Japanese doctors at international conferences: Why the worry? *Journal of Medical English Education, 12*(3), 47–55.

Hamada, A., Iso, T., Kojima, M., Aizawa, K., Hoshino, Y., Sato, K., Sato, R., Junko, C., & Yamauchi, Y. (2021). Development of a vocabulary size test for Japanese EFL learners using the new JACET list of 8,000 basic words. *JACET Journal, 65*, 23–45.

Hitosugi, M., Fukuzawa, Y., Ado, C., Mori, S., Minton, T., & Langham, C. (2016). A novel textbook based on JASMEE’s medical English education guidelines. *Journal of Medical English Education, 15*(3), 88–89.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21.

Hyland, K., & Jiang, K. (2018). Academic lexical bundles: How are they changing? *International Journal of Corpus Linguistics, 23*(4), 383–407.

ICMJE. (2024). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. Retrieved from https://www.icmje.org/icmje-recommendations.pdf

Imao, Y. (2023). CasualConc (Version 3.0.6). [Computer software]. Retrieved from https://sites.google.com/site/casualconcj

JACET Committee for Revision of the JACET Wordlist. (2003). *JACET list of 8000 basic words*. JACET.

JACET Special Committee for Revision of the JACET Wordlist. (2016). *The new JACET list of 8000 basic words*. Kirihara Shoten.

Jalali, Z. S., Moini, M. R., & Arani, M. A. (2015). Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. *International Journal of Information Science and Management, 13*(1), 51–69.

Japan Society for Medical English Education Guidelines Committee. (2015). Medical English education guidelines corresponding to the global standards for quality improvement, basic medical education: Japanese specifications. Retrieved from https://jasmee.jp/wp-content/uploads/2019/12/Guidelines_E.pdf

Jego, E. H. (2012). Corpus analysis demonstrates that scientific writing uses the structure “<Disease> was diagnosed . . .” more than “<Person> was diagnosed . . . .” *The Journal of Medical English Education, 11*(3), 50–58.

Luo, N., & Hyland, K. (2019). “I won’t publish in Chinese now”: Publishing, translation and the non-English speaking academic. *English for* Ac*ademic Purposes, 39*, 37–47.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes, 19*(1), 63–86.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics*. Cambridge University Press.

McLean, S., Hogg, N., & Rush, T. (2014). Vocabulary size of Japanese university students: Preliminary results from JALT sponsored research. *The Language Teacher, 38*(3), 34–37.

MEXT. (2014). English education reform plan corresponding to globalization. Retrieved from http://www.mext.go.jp/en/news/topics/detail/__icsFiles/afieldfile/2014/01/23/1343591_1.pdf

MEXT. (2023a). List of universities with a medical school (2023). Retrieved from https://www.mext.go.jp/content/20231010-mxt_igaku-100001063_1.pdf

MEXT. (2023b). Increase in the admission capacity of medical schools in academic year 2024. Retrieved from https://www.mext.go.jp/kaigisiryo/content/000245144.pdf

Millar, N., Budgell, B., & Fuller, K. (2012). 'Use the active voice whenever possible': The impact of style guidelines in medical journals. *Applied Linguistics, 34*(4), 393–414.

Millar, N., Salager-Meyer, F., & Budgell, B. (2019). It is important to reinforce the importance of .": 'Hype' in reports of randomized controlled trials. *English for Specific Purposes, 54,* 139–151.

Mizumoto, A. (2015). Corpus-based analysis of lexical bundles: Its potential applications in English language teaching. *Journal of the Japan Society for Speech Sciences, 16*, 30–34.

Mizumoto, A., Pinchbeck, G. G., & McLean, S. (2021). Comparisons of word lists on New Word Level Checker. *Vocabulary Learning and Instruction, 10*(2), 30–41.

Nakano, M., Miyazaki, Y., Fujieda, M., Asano, M., Noguchi, J., Ishikawa, Y., & Wakasa, T. (2021, May 22). Prototype of a medical English education support system using Japanese-English bilingual expressions in medical abstracts [Paper presentation in Japanese]. 96th Annual Meeting of the Chubu Branch of LET, Online conference.

Nakata, T. (2022). Tango no gakushu [Vocabulary learning]. In T. Nakata & Y. Suzuki (Eds.), *Eigo gakushu no kagaku* [The science of English learning] (pp. 13–30). Kenkyusha.

Nankodo. (2024). *The New England Journal of Medicine*. Retrieved from https://www.nejm.jp

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63,* 59–82.

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, I. S. P., & Macalister, J. (2021). *Teaching ESL/EFL reading and writing*. Routledge.

Nesi, H. (2013). ESP and corpus studies. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 407–426). Wiley Blackwell.

Nwogu, K. N. (1997). The medical research paper. *English for Specific Purposes, 16*(2)*,* 119–138.

Ogawa, R. (2014). Daigaku yakugakubu ni okeru kyoiku [Education in university pharmacy schools]. *The Japanese Journal of Applied Therapeutics, 6*(1), 41–46.

Oshimi, T. (2022, September 1). Igakusei no TOEFL ITP taisaku. Dr. Oshimi Medical English

Cafe #42 https://www.icrip.jp/eigocafe/2022/09/01/menu_42

Quero, B., & Coxhead, A. (2018). Using a corpus-based approach to select medical vocabulary for an ESP course: The case for high-frequency vocabulary. In Y. Kirkgöz & K. Dikilitaş (Eds.), *Key issues in English for specific purposes in higher education* (pp. 51–75). Springer.

Renouf, A., & Sinclair, J. M. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128–143). Longman.

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE, 16*(8), e0254937.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363.

Shimizu, M. (2019). Japanese medical students' reading of English academic papers and an evaluation of their ability to put grammatical knowledge to practical use. *Journal of Medical English Education, 18*(3), 82–86.

Shoecraft, K., Massa, H., & Kenway, L. (2024). Translanguaging pedagogies: Using an action research approach to support English as an Additional Language (EAL) students in a first-year undergraduate anatomy course. *Journal of English for Academic Purposes, 68*, 101357.

Simpson, A. (2022). Medical students' English needs and curricular developments. *Journal of Medical English Education, 21*(3), 94–100.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Languages, 10*(1), 65–108.

Swales, J. M. (1990). *Genre analysis.* Cambridge University Press.

Tabata, T. (2012). Dickens to Collins no kyocho sakuhin heno buntai tokeigakuteki approach [Stylometry of co-authorship: Dickens and Collins]. *IPSJ SIG Technical Report, 2012-CH-93*(3), 1–7.

Tamamaki, K., & Fujieda, K. (1998). A survey on the knowledge of medicine-related English vocabulary of Japanese medical and nursing students. *Bulletin of Liberal Arts, 18,* 87–97.

Tang, C., & Liu, Y. (2019). Construction and proposal of an academic word list for medical professionals at different developmental stages. *Journal of Medical English Education, 18*(1), 13–20.

Terauchi, H. (2016). English language education in the global world: Overview of JACET's history and challenges for its next step. *JACET International Convention Selected Papers, 3,* 2–25.

The New England Journal of Medicine. (2024a). Nihon kokunai ban. https://nejm.jp

The New England Journal of Medicine. (2024b). Article types. https://www.nejm.org/author-center

Thorndike, E. L. (1921). *The teacher's word book.* Teachers College, Columbia University.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* Teachers

College Press.

Tribble, C. (2017). ELFA vs. genre: A new paradigm war in EAP writing instruction? *Journal of English for Academic Purposes, 25*, 30–44.

Turnbull, B. (2018). Is there a potential for a translanguaging approach to English education in Japan? Perspectives of tertiary learners and teachers. *JALT Journal, 40*(2), 101–134.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442–458.

West, M. (1953). *A general service list of English words*. Longman.

Willey, I., Suzuki, H., & McCrohan, G. (2019). Piloting in-service English courses for medical professionals. *Journal of Medical English Education, 18*(3), 72–78.

Zheng, Z., & Drybrough, A. G. (2023). Translanguaging in the academic writing process. *Journal of English for Academic Purposes, 65*, 101269.

（Motoko Asano, Osaka Medical and Pharmaceutical University:
motoko.asano@ompu.ac.jp）
（Miho Fujieda, Osaka Medical and Pharmaceutical University:
miho.fujieda@ompu.ac.jp）

「論文」

明示的遂行文の遂行動詞への進行形の進出に関する予備的考察*

山﨑　　聡

Abstract

In principle, the simple present tense is used for performative verbs in explicit performative sentences such as *I ask you*, *I warn you*, and *I dedicate*; however, some scholars have pointed out that the progressive form may be used instead of the simple present tense. This paper aims to provide a both diachronic and synchronic overview of the use of the progressive form in performative verbs, mainly focusing on the extent of the inroad of the progressive into different types of performative verbs, primarily based on the Corpus of Historical American English. This study found that the use of the progressive with performative verbs is a 20th-century phenomenon; while a limited number of performative verbs have been employed in the progressive form with a moderate frequency relative to the corresponding simple present, recent decades have witnessed increased relative frequencies in more frequent performative verbs and some extension to new verbs.

The findings of this study largely align with those of De Wit et al. (2018; 2020) concerning the types of performative verbs commonly associated with the progressive form. However, it also identified several contradictory facts and extensions their studies overlooked. This paper argues that while the "extravagance" associated with the progressive form tends to be exploited with informal and colloquial performative verbs, other factors, including individual verb characteristics and additional features of the progressive form, contribute to the usage of progressive performative verbs.

1．はじめに

I beg you，I warn you，I deny，I dedicate のような英語の明示的遂行文は，そ

の主語に一人称のI，その遂行動詞には単純現在形を用いるのが典型とされる（Searle, 1989: 537; Huang, 2014²: 122など）。しかし，遂行動詞に単純現在形の代わりに現在進行形が用いられることが一部に指摘されてきた（以下，用例中の太字は筆者による）。

（1）a. I **am asking** you to do this for me, Henry, I **am asking** you to do it for me and Cynthia and the children. （Searle, 1989: 537）
b. Don’t come too close. I warn you/**I’m warning** you.
c. We propose/We **are proposing** a compromise.
（（b），（c）–Eastwood, 1994: 16）
d. You **are being discharged** on the grounds of severe temperamental unsuitability for service in the Royal Navy. （Thomas, 1995: 45）

用例（1c）と（1d）では，主語もI以外が用いられているが，ここでのポイントは，（1）のいずれにおいても遂行動詞に単純現在形の代わりに進行形が用いられている点である。Searle（1989）は遂行文の動詞に単純現在形が適している理由に触れてはいるが，進行形が用いられることに関する分析は行っていない。また，Eastwood（1994）とThomas（1995）も遂行動詞に進行形が用いられることがあるという指摘に留まっている（Fraser, 1996: 173も参照）。

その中で，近年De Wit et al.（2018）は，英語の明示的遂行文に周辺的に用いられる進行形の使用について考察を行っている。彼らの調査によると，大多数の遂行動詞には単純現在形が用いられるものの，進行形は指示型（directive）を中心とした一部の遂行動詞に特によくみられ（particularly common）（p. 257），単純現在形に比べて，話し手の強調・いらだち・緩和等の通常とは異なるextravagance（以下「桁外れ感」）を伝えるとしている。そして，彼らは，指示型の遂行動詞に進行形が用いられやすく，promiseなどの行為拘束型（commissive）やapologizeなどの表現型（expressive）では進行形は皆無であると指摘している。

こうして，De Witらにより，進行形の遂行動詞への進出がおそらくはじめて具体的に分析されたが，まだ不十分と感じられる部分も存在する。中でも，彼らの考察は現代英語についての共時的分析で，遂行動詞への進行形の進出がいつ頃からみられる現象であるのかについては不明である。また，研究で述べられているのは，進行形の遂行動詞のことのみで，対応する単純現在形との競

合をはじめ，それとの関係については触れられていない。そこで，本稿はDe Witらとは異なる関心と視点から，主にCorpus of Historical American English（COHA）を用いて，進行形はいつ頃からどのような遂行動詞にどれだけ進出しているのかを中心に，進行形の遂行動詞の使用実態について主に記述的な予備的考察を行うことを目的とする。

本稿の構成は以下の通りである。2節では先行研究としてDe Wit et al.（2018; 2020）の見解を概観する。3節では調査方法を述べる。COHAを用いてのデータ収集の手順を説明し，除外例と採用例を挙げながら本稿での用例採取の方針を示す。4節では調査結果を提示し，De Wit et at.（2018）の知見を参照つつ，進行形の遂行動詞への通時・共時的な進出状況を観察し，またその考察を試みる。5節で本稿で得られた知見を簡潔に振り返り，今後の諸課題を述べる。

2. De Wit et al.（2018; 2020）の遂行動詞と進行形に関する考察

1節で触れたように，遂行動詞における進行形の使用とその進出を詳しく扱った研究は，筆者の知るところでは，De Wit et al.（2018; 2020）のみと考えられる。本節では，遂行動詞に原則単純現在形が用いられることと進行形の基本的な意味についての彼らの考え方と，遂行動詞における進行形使用に関するその観察と知見を概観する。

まず，遂行文に用いられる相（aspect）について，De Wit et al.（2018）は，一般に，発話の時点で「完全かつ瞬時同定可能な」（fully and instantly identifiable）状況を表す（相の）形式が用いられることを16の言語で論証している。「完全かつ瞬時同定可能な」とは，当該の文によって表される事態がどのようなものか，その全体像が発話の時点で認識できる状態を指し（pp. 239–243），英語ではそれを表す形式に原則単純現在形が充てられるという。遂行動詞は，その定義上，発話と瞬時同時的にその文が表す行為を遂行することから（吉良，2018: 40），話し手は遂行文が表す事態をその発話時点で認識・把握していることになる。そのため，英語の遂行動詞には単純現在形が原則用いられる。De Wit et al.（2018）は完全かつ瞬時同定可能な状況をほかにも挙げていている。例えば，状態（state），習慣的行為，スポーツの実況中継，料理の実演，歴史的現在，確定的な（公の）予定などがそれである。[1] これらの状況はいずれも単純現在形で表される。

次に，彼ら（De Wit et al., 2009; De Wit et al., 2020）の進行形の捉え方を簡潔にみておく。彼らによれば,進行形は（通言語的に）「認識的偶発性」（epistemic contingency）というスキマティックな意味をもつ。認識的偶発性とは，当該の事態が生起する必然性がない（not necessary）ことを表す。進行形で表される事態，例えばHe is talking on the phone at the moment.は，実際に生起していることではあるが，それは予期されたことではなく，必然性のない出来事という。この点，単純現在形で表される，I walk in the park on Sunday mornings.のような習慣的行為は非有界・均質的で，その事態は瞬時同定可能であることから，その意味で予期され，必然性がある。De Witらは，進行形にまつわる，話し手の事態に対する驚き，いらだちといった主観的な意味（桁外れ感）も，発話時における予測不可能，必然性のなさという進行形の認識的偶発性に起因すると主張する。

この進行形の認識的偶発性にまつわる桁外れ感は，進行形がいまだ文法化していなかった古英語から近代初期英語にかけて，進行形の目立った用法としてしばしば指摘されてきたし,[2] De Wit et al.（2020）も近代初期英語の進行形は,単純現在形に比べて桁外れ感を伴った用例が多いことを論証している。しかし,De Wit et al.（2020）は，この桁外れ感は，進行形の文法化が進んだ現代英語にも存在するとし，単純現在形の使用が典型的な，状態動詞，遂行動詞や習慣的な行為を表す文脈等において，桁外れ感を狙って，単純現在形の代わりに進行形が用いられている事例を挙げている。[3]

本稿のテーマである遂行動詞における進行形の使用については，De Wit et al.（2018）でより詳しい調査報告とその考察がなされている。それによれば,The Corpus of Contemporary American English（COCA）による調査によると，大多数（large majority）の遂行動詞では専ら単純現在形が用いられていたが，特定の種類の遂行動詞に進行形の使用が観察されたという。つまり，遂行動詞の進行形はwarn，order，requestのような指示型の遂行動詞に目立つが，thank,apologizeのような表現型とpromiseのような行為拘束型では皆無であったという。彼らは,進行相が指示型の遂行動詞に目立つのは,指示型の動詞の中には,切羽詰まった意味をもつもの（warn，order，requestなど）もあり，それが進行形の桁外れ感と馴染むためと論じている。

3. 調査方法と用例収集

3.1 調査方法

2 節で，一部の遂行動詞に用いられる進行形は，進行形に内在する桁外れ感が現代英語にも残存している現われという De Wit et al.（2018; 2020）の見解を紹介したが，本稿では，進行形がいつ頃から遂行動詞に進出しているかについての関心から，アメリカ英語の通時的な変化を追える Corpus of Historical American English（COHA）を主に用いて調査を行った。Wierzbicka（1987）などを参考に，異なるクラスの遂行動詞の中で比較的頻度が高いと目される 49 の遂行動詞を選定して，まずは 1900 年代，1940 年代，1980 年代，2010 年代における進行形の遂行動詞と, 主に進行形の遂行動詞がみられるものについて，対応する単純現在形の遂行動詞の出現頻度を調べた。これにより，それぞれの遂行動詞の単純現在形に対する進行形の進出の度合いを，いわば定点観測的につかめる。ただし，40 年ごとのデータ取りであるので，その間のこと，あるいは 1900 年以前のことは，不明である。そこで，第 2 段階で定点を加えた全期間を対象に検索を行い，異なるクラスの遂行動詞への進行形の進出の全体像が把握できるようにした（図 1 参照）。

3. 2 用例収集

進行形の遂行文をコーパスで収集するに当たっては，慎重に該当例を拾い上げる必要がある。それが遂行文として用いられているのか，そうではなく，記述的な（descriptive）用法，現在進行中の動作を表す相的な（aspectual）用法，予め決められた予定や近接未来を表すのか，あるいはいわゆる行為解説（interpretative）用法であるのか（これらの諸例は考察の対象外）の区別が重要である。コーパスの拡張文脈の参照のみでは不十分な場合には，（利用可能であれば）小説からの用例は Google Books 等で，テレビドラマや映画からの用例はその作品にて，その文脈やシーンを吟味することで該当例を拾いあげた。以下，調査から除外したタイプの用例（3.2.1）と該当例（3.2.2）を挙げながら，本稿での用例採取の方針を示す。

3.2.1 除外例

単純現在形・進行形の遂行動詞全般について次の 3 つのタイプの用例は除外した。つまり，（i）記述的な用法，（ii）相的な用法，（iii）行為解説用法の用

例である。また，(i)～(iii) のいずれかの用法と遂行動詞とであいまいと考えられるもの（iv）も除外した。順にみてゆく。

(i) 記述的な用法

用例（2）は，「いつ Mikal に会えるのか」と尋ねたのに対して，相手は高圧的な回答を与えている場面である。ここでは，Mikal に会うことを決めるのは自分次第と述べているのであり，動詞 deny によって Mikal に会うことを禁じる発話行為をこの場で行っているわけではなく,(2) は遂行文の用例ではない。

（2）“And when will that be?” “When I say so. I make the schedule. I give access to Mikal or **I deny** access to Mikal!”　（1980, FIC, O. S. Card, *Songmaster*）

(ii) 相的な用法

用例（3a）では，先だって「君にしてあげられることを言おう」（下線部分）とあるので，この進行形は意思未来用法であろう。(3b）も，文頭の At our next meeting があることから，発話の時点では提案はしておらず，確定的な未来用法である。

（3）a. I'll tell you what I am willing to do, though. I think the Agency has grossly underestimated you, and **I'm recommending** you for a senior case officer position. Provided you still want it.　（2012, MOV, *Safe House*）

b. At our next meeting, **I suggest** a considerable increase in the funds for the Committee.　（1944, FIC, T. Caldwell, *Final Hour*）

(iii) 行為解説用法

本稿では，Ljung（1980: 69–80）や毛利（1980: 115–131）等に従い，進行形の行為解説用法を，当該の文に先立つある発言（単純形で表される）の意図または行為の意味するところを，メタ言語的に（進行形で）解説する用法と捉える。König（1980）は，行為解説用法が典型的に現れる統語環境を指摘しているが，並列構造（parataxis）に出現する場合は，形式的な手掛かりはないので，結局文脈を読み込んで判断することになる（Smitterberg, 2005: 234f.）。Smitterberg（2005: 236f.）による 19 世紀の英語の調査では，解説の対象が，König が指摘する明示的な統語環境で使用される事例は実際には少なく，該当例の 85% は（並列構造で）文脈から判断されたという。実際，下記の（4）もそのよう

な用例かと思われる。また，米倉（2023）は，近代英語期の進行形の行為解説用法を観察して，それは先行の言動を対比的に再解釈する文脈で用いられやすいことを指摘している。用例（4a）は対比的な文脈で用いられていると考えれるが，（4b）は特に対比的とは思われない。さらに，解説の対象は言語化されないこともある（Ljung, 1980: 73; Leech, et al., 2009: 134）（(6）はその例かもしれない)。そこで，本稿では，先行研究で指摘されている生起環境に関わる知見を手掛かりにしつつも，Smitterberg らと同様に，文脈を吟味することで行為解説用法か否かを判断した。行為解説用法と判断したものを 2 例のみみる。

（4）a. “I have my plans, too,” she said. “You can forget yours. I’m not going to ask you to do any of these things. **I’m ordering** you to do them. You have no choice. . . Naturally, first, you’re never to see her again.

（1949, FIC, J. O’Hara, *Rage to Live*）

b. “Hypertension is nothing to mess with, Abigail. You’re so . . . restless. You need a break—a chance to find some peace in your life.” She cleared her throat, then her face took on that I’ve-made-up-my-mind look. “Whether you go to your aunt’s or not, **I’m insisting** you take a leave of absence.”

（2011, FIC, H. Denise, *A Cowboy’s Touch*）

まず，（4a）では，これからは言うことは依頼ではなく（下線部分），命令なのだと自らの発話の意図を解説している。（4b）も話し手自身の直前の発言（下線部分）の意図を解説していると考えられる。[4)]

さらに調査では，上記の（i)～(iii）のいずれの用法であるのか判断が難しい事例もみられたが，これらはいずれにしても遂行動詞の用例ではないので，除外した。1 例だけ例示する。（5）では, I’m asking you to . . . は「帰りたいので，ディナーのことは忘れてとお願いしているのです」と，forget it の発言の意図を相手に確認させる行為解説用法と考えられる。しかし，当該の発言に先立ち forget it と数度にわたり頼んでいるので，相的用法でもあるかもしれない。

（5）Well, listen, I really appreciate the gesture but I want to go home. Why not dinner? Just forget it. The pasta will be al dente in two minutes 17 seconds. Forget it, Lois. Where’s your dinner guest? Forget it. She forget it? She didn’t forget it. **I’m asking you to** forget it. Very well, I will erase it from memory. You do that.

（1984, MOV, *Runaway*）

(iv) 遂行動詞とそのほかの用法とであいまいなもの

(i)～(iii) のいずれかの用法と遂行動詞とであいまいと考えられる用例もみられた。このタイプの用例も除外したが, 1 例だけみておく。Smitterberg（2005: 234）は，書簡中の I am writing . . . について，解釈の対象が明示的に述べられていない場合でも when I produce this letter のような節を補うこともでき，行為解説用法とれるかもしれないという。しかし，その時点で手紙を書いていることから単に相的な意味にもとれ，I am writing. . . はあいまいであろうと述べている。用例（6）の asking はその時点で瞬時同時的に依頼をするので，writing とは異なり，吉良（2018）の言う進行形の「前段階」が asking には存在しないことから相的な読みはなく,[5] 遂行動詞の進行形と考えられる。しかし，当該の文も投稿者の執筆意図を伝える（when I produce this letter, I am asking . . .）行為解説用法と取れなくもないだろう。 調査では，（6）のように，遂行動詞以外の解釈も可能で，いずれともとれる用例は除外した。

（6）I read the letters in the Post and think they bring certain problems before the public, so **I'm asking you to** print mine as soon as possible and bring the subject of tinnitus into the public awareness.

（1986, MAG, *Saturday Evening Post*, MEDICAL MAILBOX）

3.2.2 該当例

これまで調査対象から除外したタイプについて述べてきた。以下では該当例を示す。用例は Searle（1975）による遂行動詞のタイプ別に，紙幅の関係でより頻度の高い進行形の用例を中心に提示する。まず，断言型の用例（7）をみてみよう。（7a）の I'm telling you は字義通りの意味が希薄化して，ホスト文の命題内容を強調する定型句と考えてよいだろう。[6]（7b）は，ゴルフに必要なのは練習量で，ゴルフには生まれながらの才能というものは存在しないという主張を展開する文章からである。この I'm saying は，I'm telling you と同様に，ホスト文の命題内容を強調しているであろう。（7c）は，うちの芝生の庭にヒアリを蒔いたのはあなたね，とその証拠を見つけた Peggy が隣家の Dale に問い詰めると，彼はそれを否定する場面である。その場で瞬時同時的にその訴えを否定しているので，遂行文である。また，命題の真偽に関する言明であるので，

断言型と判断される。

（7）a. This is total authenticity. **I’m telling you**, Ellie, not one single thing is missing.
（2015, FIC, R. R. Cooper, *Tunneling*）

b. If you think I’m crazy, you’re going to think I’m crazier. **I’m saying** that you have as much innate golf talent as Tiger Woods. That is, you came into this world with the same inborn ability to play golf that he did.
（2013, MAG, *Golf Magazine*）

c. Peggy: How could you do it? How could you plant fire ants on our lawn?
Dale: **I’m denying** that. That’s my position.
（1997, TV, *King of the Hill*, Season (S) 1）

次に，指示型の遂行動詞の用例を提示する。まず，比較的用例数が多かった動詞の用例を（8）にみる。（8a）では重症のPTSDを病むLukeが，両親も認識できなくなり，二人をクロゼットに閉じ込め銃口を向けている。父親が「頼むから，こんなことは止めてくれ」と懇願している場面である。（JennaはLukeの妻）その場で瞬時同時的に懇願していて，「前段階」は存在せず，遂行動詞のはずである。また，懇願しているのは自明で，その意図を自分で解説する必要はなく，行為解説用法とは考えにくい。（8b）は，容疑者の少年たちを擁護する神父が，「彼らを警察署から釈放しないとマスコミを呼ぶぞ」とやって来たのに対して，警官が「手掛かりが見つかったので，もう少し時間をくれないか」と神父に頼んでいる場面である。この発言はその場でなされた瞬時同時的依頼の発話行為ととるのが妥当だろう。（8c）は，法廷での弁護士と裁判官とのやり取りで，ここで弁護士は裁判官に対して，「依頼人の自己誓約による釈放を要求します」とこの場で瞬時同時的に要請している。（8d）もこの場で緊急避難を命じている場面で，遂行動詞にちがいない。（8e）は，自分の娘Carolineがバンパイアであることに気づいた刑事でもある母親が，Carolineを連行する。そこにCarolineの友人のBonnieが現れ，彼女に対してCarolineの母親が「この件に関わらないように」と警告している場面である。Bonnieを見かけたその場でそう警告しているので遂行文である。

（8）a. Luke: What did you do to Jenna?
Lukeの父親：Luke, please, stop this, **I’m begging you**, your mother, her

heart is weak. （2011, TV, *Criminal Minds*, S7）

b. 神父：Sergeant Voight, you've had the [suspected] boys for 24 hours. If you don't release them, I'm gonna bring the press up here. 警官：Father, we're on to something. **I'm asking you** to give us a little more time. Please. （2017, TV, *Chicago P.D.*, S4）

c. 裁判官：To the charges of aggravated mayhem, kidnapping, and false imprisonment, how do you plead?
弁護士：Not guilty. And **I'm requesting** that my clients be released on their own recognizance. （2011, TV, *Drop Dead Diva*, S3）

d. ". . . And we all agree the gravity shifts are beyond anything the safety systems were designed for. With regret **I am ordering** an immediate evacuation." （2009, FIC, P. F. Hamilton, *The Temporal Void*）

e. Caroline の母親：I'd say the vampire I've been looking for is my own daughter. Bonnie:［Caroline が連行されるのに気付いて］Caroline? Caroline: Bonnie, find Damon. Caroline の母親：**I'm warning you**, Bonnie. Stay out of this. （2017, TV, *The Vampire Diaries*, S8）

次に，そのほかの指示型の用例（9）を検討しよう。(9a）は，（元）町長は人間ではなく悪魔あると住民は分かったので，新しく町長に選ばれたという男（Gibson）が，彼に「家に帰れ，できればここから出て行ってくれ」と述べる場面である。その場でそう忠告しているのであるから，遂行動詞ととるのが最も自然である。(9b)では，開店準備中のレストランに投資を考えている男性が，レストランの中身について話し合う会合にて提案を行っている。(9c）は Evans 少年にとって最適な受け入れ先をこの場で推奨している場面であるので，この recommending は遂行動詞の用法と判断される。(9d）は，集会の冒頭で，Cinch が Thank you for coming here. と述べた後に「手短に言うと，自然保護区の維持管理をわれわれの組織に移管することを提案します」と述べている部分からである。冒頭での発言であるので，ある発言の意図を解説しているというよりも，発話と瞬時同時的に提案の発話行為を行っているととるのが妥当であろう。最後に（9e）は，極めて難しい手術に立ち会い，何時間も処置できずにいる Shepherd 医師に上司の医師が手術中断を求めている場面である。当該の箇所は So, I am demanding . . . と So のつなぎ語があるので，その場の状況を受けて要請の発話行為を行っていると考えるのが自然であろう。

（9）a. Gibson: . . . You tricked the fine men and women of Bon Temps into thinking they were voting for a human being when the truth is they were voting for the Devil. ある住民：You tell him, Mayor Gibson. Gibson: The people have spoken. I'm the mayor now. And as mayor, **I'm advising** you to go on home, or better yet, leave Bon Temps for good.（2014, TV, *True Blood*, S4）

b. How about this idea? I feel that you can never get a waiter's attention. So, **I'm suggesting** that every table should have a bell on it.（2002, TV, *Curb Your Enthusiasm*, S3）

c. Evans is a war of the state, and it is this court's responsibility to decide where he would be best placed. As a representative of Bridgepoint Social Services, **I am recommending** that the Border home is the best environment for Evans to grow up in.（2011, MOV, *Beyond Acceptance*）

d. "Thank you for coming here," Cinch started saying, first making sure that everyone had a seat and a cold soft drink in front of them. Having heard about the government's emphasis on its need for transparency, he said, "To be quick, **I'm proposing** that you transfer the maintenance of this natural preserve to our organization."（2004, FIC, A. Kuo, *Free Kick*）

e. Dr. Shepherd. The rate of infection for this patient is increasing every second you keep him open. Not to mention the thousands of dollars you are wasting standing here doing nothing. So **I am demanding** that you close this man up. Close him up and relinquish the OR.（2009, TV, *Grey's Anatomy*, S6）

調査では，表現型と行為拘束型の用例は皆無であったため，最後に宣言型の諸例をみておく。まず，（10a）は小説の献辞からであるが，まさに献辞あることから I'm dedicating は遂行動詞にちがいない。（10b）は危険人物が脱走したので，要員に武器を与え，厳戒態勢を取れと述べている場面である。その場でそう述べて緊急事態を宣言する遂行動詞といえる。（10c）では，大統領がこの辞任宣言に続く手短な挨拶の後に実際に退任している。（10d）では，当該の場面の直前に殺人現場に手を付けるな，と少佐が述べている。この少佐の発言を聞いて，大佐が彼に「お前をスコットの弁護人に任命する」と述べている。少し後で I've appointed you counsel.（下線部分）と完了形で述べていることは，当該の文が遂行文であることの証左であろう。最後に（10e）は，離れたところからおもちゃのカエルを器に投げ入れるゲームで，ウサギのぬいぐるみを獲

得した男子が，ゲームを主催している女子に名前を尋ね，「そうしたら，このウサギを Hannah と名付けるよ。」と言って名づけ，それを与えようとする場面である。I'm naming . . . という発言と同時に名付けているので，遂行動詞にほかならない。[7]

（10）a. Author's Note: This story is for my Aunt Michelle, who recently passed due to aneurysm and stroke. **I'm dedicating** this short story to her.
（2019, FIC, *See You Again*）

b. "Mister Jeffers," he said abruptly, "break out the stun guns. Issue one to each officer and one to each chief non-com. Until we get this straightened out, **I'm declaring** a state of emergency."　（1962, FIC, G. Randall, *Unwise Child*）

c. Effective immediately, **I am resigning** the presidency of the United States.
（COCA, 2014, TV, *House of Cards*, S2）

d. 大佐：**I'm appointing** you counsel for Lieutenant Scott. 少佐：Sir, I'm not a lawyer. 大佐 : You sounded like one a minute ago. 少佐：I could be a material witness. I mean, I heard the lieutenant going out. 大佐：The lieutenant needs our help. <u>I've appointed you counsel.</u> Understood? 少佐：Yes, sir.　（COCA, 2002, MOV, *Hart's War*）

e. 男子：I'd like the pink bunny, please. What's your name?　女子：Hannah. 男子：Well, **I'm naming** my bunny after you, Hannah.
（2006, TV, *Veronica Mars*, S2）

4. 調査結果と考察

4.1 調査結果

本節では，2 節と 3 節で述べた方針で収集した該当例の集計結果を提示し，De Wit et al.（2018）の知見に言及しながら，進行形の遂行動詞への進出状況を観察する。まず，（単純現在形に対して）進行形の遂行動詞が比較的多くみられたものと一部単純現在形の件数の多いものについて，1900 年代から 2000/2010 年代の間の 4 つの時期における単純現在形と進行形の出現頻度を示したものが表 1～表 3 である。[8] 表では，動詞は Searle（1975）による遂行動詞のクラス別に提示してある。原則，概ね 40 年ごとのデータを採取したが，I tell you/I'm telling you は件数が多いことから, 3 つの時期からデータを採取した。

また, I’m saying では, I say の検索結果が多い（例えば 2010 年代だけで 2,837 件）ため，COHA 全体の件数は I’m saying だけの件数をまず示し，I say との比較は TV/MOV のサブコーパスにおける件数を右側に提示した。粗頻度の直後のカッコ内の数値は，それぞれ 100 万語当たりの調整頻度を表し，比較的高頻度のものには，その頻度に応じた濃淡の網掛けを施した。[9] なお，進行形の件数は I’m の縮約形と I am の用例を含むが，表では縮約形で代表させている。全期間を通して該当例が希少な宣言型の遂行動詞（表 3）については，COCA による調査結果（C. と表記）も併記した。

表 1. 単純現在形と進行形の遂行動詞の通時的出現件数〈断言型〉[10]

	1900s	1960s	2010s
I tell you	430 [19.57]	496 [17.03]	75 [2.12]
I'm telling you	5 [0.2]	**101 [3.47]**	**162 [4.57]**

	1900s	1940s	1980s	2010s
I'm saying	0	3 [0.11]	2 [0.07]	5 [0.14]

[TV/MOV]	1940s	2010s
I say	42 [14.54]	15 [2.96]
I'm saying	3 [1.04]	2 [0.40]

表 2. 単純現在形と進行形の遂行動詞の通時的出現件数〈指示型〉

	1900s	1940s	1980s	2010s		1900s	1940s	1980s	2010s
I ask*	50 [2.28]	49 [1.79]	46 [1.54]	26 [0.73]	I beg **	188 [8.55]	119 [4.34]	108 [3.62]	51 [1.44]
I'm asking	0	7 [0.26]	9 [0.30]	13 [0.37]	I'm begging	0	1 [0.04]	12 [0.40]	34 [0.96]

	1900s	1940s	1980s	2000s+10s		1900s	1940s	1980s	2000s+10s
I request	3 [0.14]	13 [0.47]	11 [0.37]	17 [0.24]	I order	2 [0.09]	9 [0.33]	15 [0.50]	34 [0.48]
I'm requesting	0	1 [0.04]	1 [0.03]	6 [0.09]	I'm ordering	0	0	9 [0.30]	17 [0.24]

	1900s	1940s	1980s	2010s		1900s	1940s	1980s	2000s+10s
I warn	57 [2.59]	79 [2.88]	45 [1.51]	17 [0.48]	I advise	36 [1.64]	39 [1.42]	29 [0.97]	34 [0.48]
I'm warning	0	27 [0.99]	47 [1.54]	23 [0.65]	I'm advising	0	0	0	2 [0.03]

	1900s	1940s	1980s	2010s		1900s	1940s	1980s	2010s
I suggest	29 [1.32]	176 [6.42]	217 [7.27]	176 [4.96]	I propose	58 [2.64]	39 [1.42]	30 [1.00]	30 [0.85]
I'm suggesting	0	1 [0.04]	0	2 [0.06]	I'm proposing	0	0	1 [0.03]	0

	1900s	1940s	1980s	2010s		1900s	1940s	1980s	2010s
I recommend	18 [0.82]	18 [0.66]	43 [1.44]	52 [1.47]	I insist	36 [1.64]	60 [2.19]	66 [2.21]	55 [1.55]
I'm recommending	0	0	0	2 [0.08]	I'm insisting	0	0	0	0

*I ask a favor (of you) は除外した

**I beg your pardon/forgiveness/leave、I beg to differ/disagree は除外した

表 3. 単純現在形と進行形の遂行動詞の通時的出現件数〈宣言型〉

	1900s	1940s	1980s	2010s	C. 2000s+10s
I dedicate	4 [0.18]	3 [0.11]	3 [0.10]	3 [0.08]	39
I'm dedicating	0	0	0	1 [0.03]	12

	1900s	1940s	1980s	2000s+10s	C. 2015-19
I declare	3 [0.14]	8 [0.29]	7 [0.23]	9 [0.13]	18
I'm declaring	0	0	1 [0.03]	2 [0.03]	3

	1900s	1940s	1980s	2000s+10s	C. 2010s
I resign	2 [0.09]	4 [0.15]	2 [0.07]	1 [0.01]	9
I'm resigning	0	1 [0.04]	1 [0.03]	4 [0.06]	5

	1900s	1940s	1980s	2010s	C. 1990s-2010s
I appoint	1 [0.05]	4 [0.15]	1 [0.03]	0	7
I'm appointing	0	0	1 [0.03]	1 [0.03]	10

	1900s	1940s	1980s	2010s	C. 2000s+10s
I name	0	0	0	2 [0.07]	9
I'm naming	0	0	0	2 [0.07]	6

一般に，明示的遂行文は現代英語では形式ばった印象を与え，日常的にはあまり用いられないとされるが（Thomas, 1995: 47–49; Grundy, 2020^{4}: 33f. など），その使用場面が限定的と考えられる宣言型の遂行動詞（表 3）を除けば，概ね通時的な減少傾向が表 1 と表 2 でも確認される。そのためここでは，遂行動詞への進行形の進出をそれぞれの単純現在形との相対的な比較で主にみてみよう。まず，表中で最も高頻度の進行形の遂行動詞は定型的な I’m telling you で，出現の時期が最も早く，2010 年代の相対的な頻度も単純現在形の 2 倍以上となっている。続いて，指示型（表 2）の「依頼」(asking，begging と requesting)，「警告」(warning)，「命令」(ordering) を表す遂行動詞の相対的な頻度が高い。このうち asking と warning は 1940 年代の比較的早い時期から一定程度見られるが，全体的に概ね現在に近づくほど相対的頻度を上げていて，warning では 1980 年代より単純現在形を上回っている。宣言型の遂行動詞(表 3)は，その動詞の意味からコーパスでの出現頻度こそ低いが，進行形は概ね 1980 年以降散見され，その近年の進出の様子は，それぞれ右欄に併記した COCA における単純現在形との競合にうかがえる。

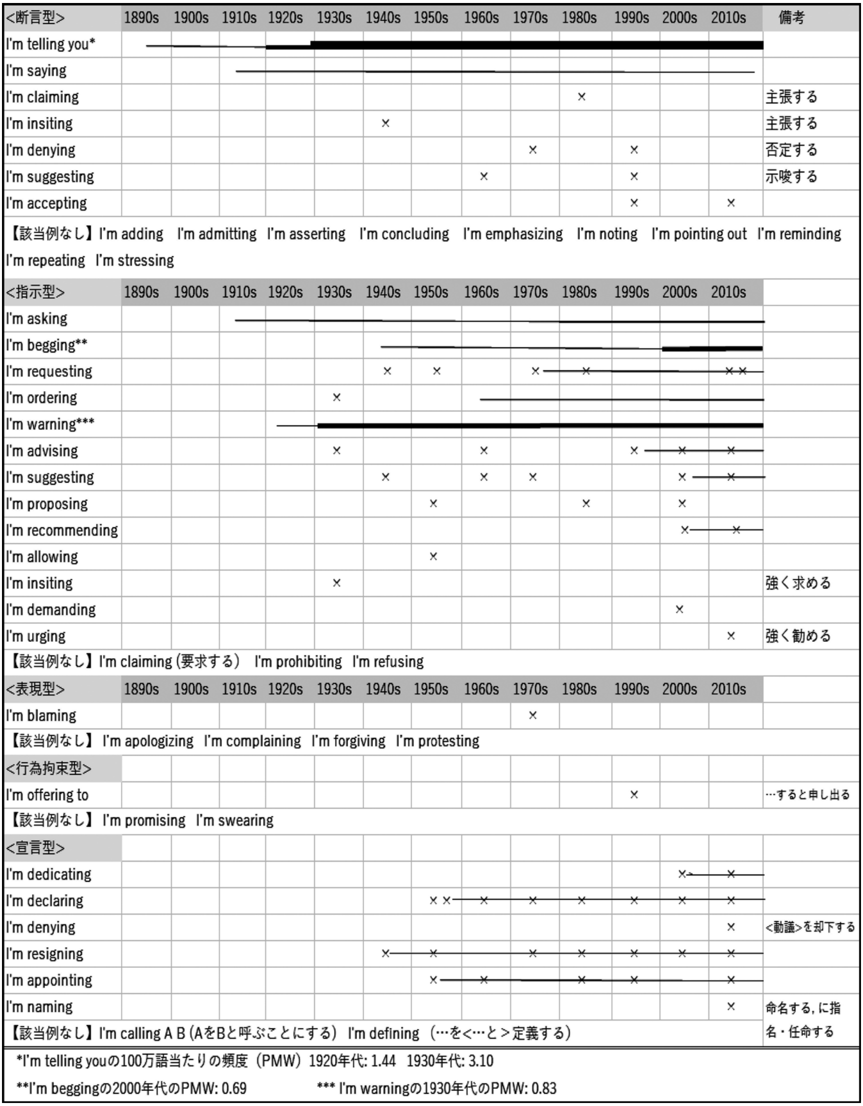
最後に I’m saying（表 1）について触れておこう。I’m saying は I’m telling you と同じ発話動詞であるが，ずっと頻度が低い。これには元となる単純現在形の

命題を強調するI sayが，遂行動詞の進行形が散見されるようになる20世紀半ばには衰退傾向にあり，古風になりつつあったことによると推察される。[11)]

さて，ここまで主に進行形が比較的多くみられた遂行動詞を選定して，その進出の時期と度合いを単純現在形との比較で定点観測的にみてきた。しかし，これだけでは1900年以前のことや「定点」以外のことは不明である。そこで，表1〜3で触れていないものを含めて，49の遂行動詞の進行形のCOHA全体における出現状況を図1に示す。ただし，I'm telling you，I'm begging you，I'm warning you，I'm asking you，I'm orderingとI'm sayingのより高頻度の6つの動詞では, 初出例を含めた初期や定点間の空白を埋めるのに必要な検索に留めた。ここでも遂行動詞はクラス別に提示し，be動詞は縮約形で代表させている。より高頻度の遂行動詞は，おおよそその時期の頻度に見合った太い実線で，低頻度のものは細い実線で示した。該当例が散見される程度の動詞では，それが1例〜3例確認された年代にバツ（×）印を，5例〜6例が確認された年代にはバツ印を2つ振り，バツ印に一定の連続性がみられるものには細い実線を施した。「備考」欄には，当該クラスでの動詞の意味を記載している。

まず, 図1では, 該当例が全く確認されなかったもの（図中の【該当例なし】）も少なくなく，継続して進行形の使用がみられる動詞は1/5程度に限られていている。これは，大多数の遂行動詞では専ら単純現在形が用いられるという，De Wit et al.（2018）の観察と大方一致する。De Witらは，進行形の遂行動詞は指示型の遂行動詞に目立ち，表現型と行為拘束型では皆無であると報告しているが（p. 257f.），このことも確認される。ただ図1では，断言型においても，定型的なI'm telling youとI'm sayingを除けば，進行形の進出は極めて鈍いといえる。一方で，図1では進行形は近年より多くの遂行動詞に散見され，また，単純現在形に対して進行形の相対的頻度が高い動詞では，その頻度が上昇していることを表1〜表3でみた。このことから，進行形は少しずつ遂行動詞に進出しつつあるとみてよいだろう。

次に，進行形が最も広くみられた指示型の動詞群についてみる。De Wit et al.（2018: 258）は，ある種の遂行動詞は切羽詰まった（urgent）意味をもち，このことが進行形に内在する強調や緩和の桁外れ感と親和性があると述べている。図1でも切羽詰まった，強い意味をもつと考えられるask，beg，request，order，warnに進行形が目立つ。そして，より意味が弱いと感じられるadvise，suggest，propose，recommend，allowでは進行形は少ない。しかし，強い意味をもつと考えられるinsist，demand，urgeでは進行形は極めてまれで，claim,



<断言型>	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s	備考
I'm telling you*	—	—	—	━	━	━	━	━	━	━	━	━	━	
I'm saying			—	—	—	—	—	—	—	—	—	—	—	
I'm claiming										×				主張する
I'm insiting						×								主張する
I'm denying									×		×			否定する
I'm suggesting								×			×			示唆する
I'm accepting											×		×	
【該当例なし】I'm adding　I'm admitting　I'm asserting　I'm concluding　I'm emphasizing　I'm noting　I'm pointing out　I'm reminding　I'm repeating　I'm stressing														
<指示型>	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s	
I'm asking			—	—	—	—	—	—	—	—	—	—	—	
I'm begging**						—	—	—	—	—	—	━	━	
I'm requesting						×	×		×—	×	—	—	××	
I'm ordering					×			—	—	—	—	—	—	
I'm warning***				—	━	━	━	━	━	━	━	━	━	
I'm advising					×			×			×—	×	×	
I'm suggesting						×		×	×			×—	×	
I'm proposing							×			×		×		
I'm recommending												×—	×	
I'm allowing							×							
I'm insiting					×									強く求める
I'm demanding												×		
I'm urging													×	強く勧める
【該当例なし】I'm claiming (要求する)　I'm prohibiting　I'm refusing														
<表現型>	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s	
I'm blaming									×					
【該当例なし】I'm apologizing　I'm complaining　I'm forgiving　I'm protesting														
<行為拘束型>														
I'm offering to											×			…すると申し出る
【該当例なし】I'm promising　I'm swearing														
<宣言型>														
I'm dedicating												×—	×	
I'm declaring							××	×	×	×	×	×	×	
I'm denying													×	<動議>を却下する
I'm resigning						×—	×	—	×	×	×	×	×	
I'm appointing							×—	×	—	×	×	—	×	
I'm naming													×	命名する，に指名・任命する
【該当例なし】I'm calling A B (AをBと呼ぶことにする)　I'm defining（…を<…と>定義する）														

*I'm telling youの100万語当たりの頻度（PMW）1920年代: 1.44　1930年代: 3.10

I'm beggingの2000年代のPMW: 0.69　　* I'm warningの1930年代のPMW: 0.83

図 1. 各種進行形の遂行動詞の通時的出現分布［COHA］

refuse，prohibit では観察されず，強い意味をもつ動詞が必ずしも進行形を受け入れやすいとは一概にいえないことが分かった。

最後に，宣言型の動詞群について触れておく。De Wit らは，進行形の桁外れ感は「I’m dedicating のような半ばイディオム的な表現を生んでいる（leading to more or less idiomatic expressions）」（p. 258）と述べている。しかし，図 1 では進行形はほかの宣言型の動詞にも比較的進出しつつあることが分かる。

4.2 考察

4.1 でみた異なるクラスの遂行動詞への進行形の進出状況はどのように説明されるだろうか。部分的な説明に留まるが，考察を試みる。簡潔に述べれば，De Wit et al.（2018; 2020）は遂行動詞の進行形は桁外れ感を狙って用いられるとするが，遂行動詞の進行形の使用を促す要因はそれだけではなく，ほかにも複数あると考えられる。その要因の中には，遂行動詞にまつわる「カジュアル感」や「口語的な語感」（以下，単にカジュアル感）があり，桁外れ感を狙った進行形はこのタイプの動詞で利用されやすいことがまずあると考えられる。進行形とカジュアル感，話し言葉との親和性はしばしば指摘されるところであるが（例えば Biber, et al., 1999: 461, 471; Williams, 2007: 78），形式ばった遂行文においては進行形はカジュアル感のある動詞を好むのかもしれない。また，形式ばった語感の動詞よりカジュアルな動詞の方が，話し手の桁外れ感を乗せ

表 4. 各種遂行動詞の COCA の TV/M（と FIC）のジャンルでの相対的出現状況

<断言型>	tell	say	claim	insist	deny	accept	add	admit	
	◎	△	×	×	▲	▲	×	△	
	assert	conlude	emphasize	note	point out	remind	repeat	stress	
	×	×	×	×	×	△	▲	×	
<指示型>	ask	beg	request	order	warn	advise	suggest	propose	
	○	◎	▲	△	△	▲	×	×	
	recommend	allow	insist	demand	urge	refuse	porohibit		
	×	▲	×	×	×	▲	×		
<表現型>	blame	forget	complain	protest	apologize	<行為拘束型>	offer	promise	swear
	△	◎	▲	▲	◎		×	○	◎
<宣言型>	dedicate	declare	resign	appoint	name	define			
	×	×	▲	×	△	×			

やすいこともあるかもしれない。まず，この動詞のカジュアル感の要因が，時にそのほかの要因も絡んで，遂行動詞の進行形使用にどう影響するかを吟味しよう。

それぞれの動詞のカジュアル感は主観的なものであるが，ここではそれを各動詞がマルティ・ジャンルのCOCAにて，よりフォーマルなMAG，NEWSとACADに対して，カジュアルなジャンルのTV/MOV（とFIC）で相対的にどの程度用いられるかにより，5段階で捉えてみた。具体的には，当該の遂行動詞の使用が，（1）TV/MOVとFICで際立ち，MAG，NEWS，（特に）ACADで少ないもの（◎），（2）TV/MOV（とFIC）でより多く，MAG，NEWS，（特に）ACADで少なめなもの（○），（3）ジャンル間で比較的均等で，TV/MOV（とFIC）は特に多くも少なくもないもの（△），(4) TV/MOV（とFIC）で少なめで，MAG，NEWS，（特に）ACADで多めなもの（▲），（5）TV/MOV（とFIC）で少なく，MAG，NEWS，（特に）ACADで多めの（多い）もの（×）の5つに分類した。表4は各動詞の（1）～(5）の段階を直観的な◎，○，△，▲，×で示したものである。分類は,その動詞としてのカジュアル感を捉える意図から,異なるジャンルにおける各動詞のレンマ（_vvのタグ使用）の件数に依った。いずれの段階に分類されるか微妙なものも時にあったが，各動詞のおおよそのカジュアル感の度合いを捉えることを意図している。ただし，例えば，断言型のsuggest「示唆する」は，指示型の「提案する」の用例に比べて少ないと考えられ，表には含めていないものもある。また，stressは各ジャンル任意の100例中の「強調する」の意味の用例を手作業で抽出し，その数を基にした各ジャンルの見込みの件数に依った。網掛けの動詞は，図1及び表2と3にて進行形が比較的多かったものを示す。

まず，断言型と指示型の動詞から検討しよう。網掛けのある動詞は，requestを除けば，そのほかの進行形の進出が鈍い動詞とは異なり，△，○または◎の分布となっている。ただし，進行形の使用を促進するのは，動詞のカジュアル感だけではなく，個々の動詞の事情が影響することもある。Requestは▲であるが，類義のaskとbegの影響で進行形の使用が促されていると推察される。また，warnは△で，カジュアル感は特に高くはないが，I'm telling you，特にI'm begging youと同様に，補部を従えるのではなく，*I'm warning you*, step back.のように自立的にホスト文と並列的にほぼ常に用いられる。I'm warning youは独立したユニットして認識されることで，その使用が自動的となり，このことが高頻度につながっていると考えられる。また, sayは伝統的なI sayのなごりで,

少ないながらも継続して観察されるのだろう。

これ以外の断言型と指示型の進行形の進出が鈍い動詞は，ほぼ TV/MOV（と FIC）における使用度が低く，同じ指示型の強い意味をもつ動詞（表 4 の insist ～prohibit）でも桁外れ感を狙った進行形で利用されることが少ないのは，そのカジュアル感不足にあるのかもしれない。また，断言型の accept，add，admit，conclude，note と point out では，カジュアル感不足に加えて，意味的に桁外れ感とはあまり関係がなさそうなことも，進行形が希少であることの原因であるかもしれない。ただし，図 1 と用例（7c），（9）の諸例にみるように，カジュアル感が弱い動詞でも，桁外れ感を狙った（と考えられる）進行形がまれに用いられることはあり（注 12 も参照），桁外れ感を狙った進行形はカジュアル感のある動詞で利用されるというのは「傾向」と捉えられる。

一方で，ほかのクラスの遂行動詞の多くでは，カジュアル感のあるものが進行形を受け入れやすいという傾向と矛盾する結果となっている。まず，宣言型の動詞は，近年進行形の進出が比較的よく観察されたものも含めて，TV/MOV（と FIC）での使用度が低い。また，査読者が指摘する通り，その用例も桁外れ感は関係しないもの（特に（10c）–（10e）など）がより多いと思われる。このことから，宣言型の動詞では別の理由により進行形が用いられることが多いと考えられる。このことを 3 つの動詞についてみてみよう。Williams（2007）は，法律文書では遂行動詞は単純現在形が原則としながらも，くだけた法律文書では時に進行形が用いられることがあるという。そして，*I am hereby appointing* Jean Graham as the temporary clerk for …の実例を挙げ（p. 60），ここでは当該の任命が一時的なものであるので, 進行形が用いられているという。COCA の I'm appointing の 1990 年代～2010 年代の用例を吟味すると，9 例中（10 例中 1 例は不明）4 例はこうした臨時的任命の文脈ととれそうである（（10d）参照；少佐を弁護人に任命することは通例ありえない）。同様に，I'm naming は，6 例中 4 例は臨時的な名づけ行為の感が強いと感じる（（10e）参照）。一方，I name の諸例にはそのような感はなかった。Dedicate にもこれに関連した違いがみられた。I dedicate は書籍，エッセー，建造物など制作物を捧げる事例が目立つが，I'm dedicating では，捧げるものはその場で歌う歌や試合など一過性のものが目立ち，制作物はおそらくウェブ掲載の短編（（10a））からの 1 例のみであった。[12] 宣言型の進行形は，進行形の一時性・臨時性という特性に動機づけられている部分がより大きいと考えられる。[13]

最後に，表現型と行為拘束型の動詞では TV/MOV（と FIC）の比率が高く，

カジュアル感のあるものも目立つが, 既述の通り, これらの進行形は皆無であった（図 1 参照）。現在のところ，これらのクラスの遂行動詞の進行形の使用を妨げる要因は不明で, [14] 今後の課題としたい。

まとめると，Det Wit らの指摘する桁外れ感を狙った進行形は，カジュアル感のある動詞で活用されやすいこと，一方で，個々の動詞の事情やほかの進行形の特性により進行形の使用が促されることもあり，遂行動詞の進行形使用を促す要因は複数あることを論じた。[15]

5. おわりに

本稿は遂行動詞への進行形の進出を通時・共時の両面から調査・考察を行った。De Wit et al.（2018; 2020）の知見が確認された部分もあった一方で，切羽詰まった意味をもつ遂行動詞が必ずしも進行形に馴染むわけではないこと，断言型も定型的なものを除いて進行形は希少であること，宣言型に進行形が相対的に進出しつつあることをみた。そして，遂行動詞の進行形を促す要因は，桁外れ感だけではなく，動詞のカジュアル感ほか複数あることを指摘した。

一方で，課題も残された。宣言型の遂行動詞に進行形が拡がりつつある要因はさらに究明が必要で，表現型と行為拘束型の遂行動詞では逆に進行形が避けられる理由の解明が待たれる。遂行文と行為解説用法（とそのほかの用法）の区分をどう捉えるかの問題もある。また，I'm telling you，I'm begging you と I'm warning you の自立的ユニット性（とその使用頻度との関係）については，別に論じたい。ほかにも，単純現在形と進行形の遂行文とで，主に用いられる発話行為の種類が異なる変わり種もみられた。進行形の遂行文は周辺的な現象ではあるものの，思いの外興味深い事象が存在するようである。

注

* 査読の先生方より貴重なご指摘とご意見を賜りましたこと, 厚くお礼申し上げます。なお，本稿の誤りや残された問題は筆者の責任に帰することは言うまでもない。

1. De Wit et al.（2018）によれば，状態と習慣的行為の事態は非有界で均質的であるため，また，スポーツの実況では，パス－シュートなどプレーの一連の展開はかなり予測可能であるため，話し手は当該の文の発話時にその事態を完全に認識可能という。
2. 例えば，Hübler（1998: Chap. 4），Kranich（2010: 82–88），Petré（2017）を参照。
3. 現代英語においても，進行形が強調といった主観的・感情的なニュアンスを添え

るという主張は，Hatcher（1951）や Hübler（1998: Chap. 4）などにもみられる。Hatcher（1951）は，通常は単純（現在）形が用いられるところで進行形が用いられる一連の事例の中に，I'm warning you と I'm telling you を挙げている（p. 272）。遂行動詞とは述べていないが，こうした動詞の進行形が強調の意味をもちうることをすでに指摘している。

4. 行為解説用法とそのほかの進行形の用法との切り分けは困難を伴うことが，Smitterberg（2005）と米倉（2023）で指摘されているが，それとは別に，査読者より行為解説用法と遂行動詞とを切り分けることについて根源的な問題提起をいただいた。つまり，例えば（4a）で I'm ordering は話し手の発話の意図は命令なのだと解説をしているが，それは聞き手にその意図が命令だと理解させることで，同時に発話内の力をもち，遂行文としても機能しているのではないか，というご指摘である。本研究は，遂行動詞と行為解説用法を独立したものであることを前提にしているが，今後の研究に当たってはその前提についても検討課題としたい。
5. 吉良（2018: 199–205）の言う「前段階」とはある出来事の時間的に前の領域を指し，進行形には必ずこの「前段階」が存在するという。例えば，When I visited him, he **was having** lunch. では，「ランチを食べる」行為は，「私が彼のところを訪れた」基準時より前から始まっている（「食べていた」）ことになり，「前段階」が存在する。ちなみに，発話と瞬時同時的に行為をなす遂行動詞にはこの「前段階」が存在しないために，遂行動詞は通例，進行形にならないことになる。
6. *Longman Dictionary of Contemporary English*, 5th ed. (LDOCE5)（s.v. *tell* v. 15）など複数の学習者向け英英辞典も，I'm telling you を（Spoken）Phrases あるいは Idioms として記載している。
7. 進行形と単純現在形の遂行動詞の合計 30 の用例（多くは本稿の用例とは異なる）について，2 名の英語母語話者に 2 つの形式のニュアンスの違いをアンケート調査したところ，平均で半数強のペアについて進行形の方がより強調的，深刻さを感じるといった回答（これらは桁外れ感）を得たが，特段に違いは感じられないというものもあった。また，桁外れ感は文体的・修辞的なもの（De Wit et al., 2018: 258）であるので，個々の文に対する印象には個人差も少なくなかった。
8. 集計では，近接（隣接）して用いられた用例とコーパス中の重複例は，それぞれ 1 例として数えた。
9. 網掛け対象の I'm saying の頻度は, 左側の COHA 全体での出現頻度に基づいている。本稿で調査した遂行文はほぼ TV/MOV と小説をはじめとした会話部分にみられた。
10. 表 1 は断言型の用法についての調査であるため，それ以外の用法（指示型，質問型（rogative））の用例，I say の提案や注意喚起など（LDOCE5, s.v. *say*[1]. 37）の用例は除外している。
11. TV/MOV の I say の 1940 年代の出現頻度は高いが，I tell you の TV/MOV に限った 1960 年代の出現頻度は 45.44/PMW で, 1940 年代の I say の 3 倍以上に上る。また，Yamazaki（2023: 57）のイギリス英語の戯曲と映画作品の調査によれば，命題を強調する I say の頻度は 18 世紀から 1990 年代の間にほぼ 1/9 にまで減少している。

12. De Wit et al.（2020: 504）は，単純現在形に比べて強意的な桁外れ感を伴うとする I'm dedicating の用例を挙げているが，（10a）についてインフォーマントのひとりは同様の印象を報告し，もうひとりは少しだけ強意的であるかもしれないと回答した。（10a）の進行形は桁外れ感を狙った進行形の使用例であるかもしれない。
13. ただ，これら 3 つの動詞の一時性・臨時性は感じられない用例については，何が進行形の使用を促すのかの考察が必要である。また，resign（辞職する）の用例については一時性も桁外れ感もほぼ関係ないと考えられ，その進行形を促す要因は不明である。なお，declare については，その進行形は単純現在形と異なり，（小）節でなく NP（（10b）参照）をその補部にとる傾向がみられたが，これは進行形のカジュアル感と関係があるかもしれない（注 15 参照）。
14. De Wit et al.（20218: 258）は表現型と行為拘束型の遂行動詞に進行形が用いられにくい理由をごく簡潔に示唆しているが，その説明は（切羽詰まった意味の）指示型の動詞とはどう異なるのか筆者には合点が行かない。
15. ほかにも，進行形のカジュアル感が遂行動詞の進行形使用を促すこともあるようである。インフォーマントによると，雑誌の読者欄からの用例，*I'm suggesting* vinho verde to everyone looking for a value white wine. So crisp, so clean, so citrusy, it complements just about any summer meal. Drink up!（2011, MAG, *Country Living*）では，進行形のカジュアル感がこの文の内容にふさわしいという。

参考文献

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

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

Davies, M. (2010) *The Corpus of Historical American English (COHA)*. Available online at https://www.english-corpora.org/coha/.

De Wit, A. and F. Brisard (2009) "Expressions of Epistemic Contingency in the Use of the English Present Progressive." *Papers of the Linguistic Society of Belgium* 4 (18 pages).

De Wit, A., F. Brisard and M. Meeuwis (2018) "The Epistemic Import of Aspectual Constructions: The Case of Performatives." *Language and Cognition* 10: 234–265.

De Wit, A., P. Petré and F. Brisard (2020) "Standing Out with the Progressive." *Journal of Linguistics* 56: 479–514.

Eastwood, J. (1994) *Oxford Guide to English Grammar*. Oxford: Oxford University Press.

Fraser, B. (1996) "Pragmatic Markers." *Pragmatics* 6, 1: 167–190.

Grundy, P. (2020^4) *Doing Pragmatics*. London and New York: Routledge.

Hatcher, A. G. (1951) "The Use of the Progressive Form in English: A New Approach." *Language* 27, 3: 254–280.

Huang, Y. (2014^2) *Pragmatics*. Cambridge: Cambridge University Press.

Hübler, A. (1998) *The Expressivity of Grammar: Grammatical Devices Expressing Emotion*

Across Time. Berlin: Mouton De Gruyter.

吉良文孝（2018）『ことばを彩る 1 ―テンス・アスペクト』東京：研究社.

König, E. (1980) "On the Context-dependence of the Progressive in English." In Rohrer, C. (ed.) *Time, Tense, and Quantifiers*. Tübingen: Max Niemeyer Verlag, pp. 269–291.

Kranich, S. (2010) *The Progressive in Modern English: A Corpus-based Study of Grammaticalization and Related Changes*. Amsterdam: Rodopi.

Leech, G., M. Hundt, C. Mair and N. Smith (eds.) (2009) *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Ljung, M. (1980) *Reflections on the English Progressive*. Göteborg: Acta Universitatis Gothoburgensis.

毛利可信（1980）『英語の語用論』東京：大修館書店.

Petré, P. (2017) "The Extravagant Progressive: An Experimental Corpus Study on the History of Emphatic [BE Ving]." *English Language and Linguistics* 21, 2: 227–250.

Searle, J. R. (1975) "A Classification of Illocutionary Acts." *Language in Society* 5: 1–23.

Searle, J. R. (1989) "How Performatives Work." *Linguistics and Philosophy* 12, 5: 535–558.

Smitterberg, E. (2005) *The Progressive in 19th-century English: A Process of Integration*. Amsterdam: Rodopi.

Thomas, J. (1995) *Meaning in Interaction: Introduction to Pragmatics*. London and New York: Routledge.

Wierzbicka, A. (1987) *English Speech Act Verbs: A Semantic Dictionary*. Sydney: Academic Press.

Williams, C. (2007^2) *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts*. Bern: Peter Lang.

Yamazaki, S. (2023) "Changes in *I Tell You* and Other Formulaic Explicit Assertive Performatives."『近代英語研究』39: 39–66.

米倉よう子（2023）「英語の解釈的進行形用法の発達について」近代英語協会第 40 回大会 口頭発表.

（山﨑　聡　千葉商科大学　E-mail: s2yamaza@cuc.ac.jp）

「論文」

A Study of the *NP as it is known* Expressions through Comparison with the *NP as we know it* Expressions

Yoshiaki SATO

Abstract

The aim of this paper is to elucidate the specific characteristics of *NP as it is known* expressions (e.g., *the TPP as it's known*) through corpus analysis and comparison with seemingly similar *NP as we know it* expressions (e.g., *the world as we know it*). While these two types of expressions have their surface commonalities, corpus findings reveal divergent distributional patterns for them. Specifically, the analysis revolves around (i) an association between the two types of expressions and particular situations (e.g., denoting disappearance), (ii) usage of *as*-clauses, and (iii) discourse functions. The results demonstrate that the *NP as it is known* expressions lack any significant associations with particular situations and the *as*-clauses have two important uses: naming as the main use and restricting as another less typical use. Name-*as* is used to name an entity whose linguistic expression is modified by the *as*-clause; Restrictive-as is used to restrict the scope of such an entity. Furthermore, the name-*as* clause tends to signal that the NP modified by the *as*-clause is a paraphrase of the preceding expression and this NP tends to be a primary topic in the subsequent discourse. These properties sharply contrast with those of the *NP as we know it* expressions. Overall, this study advances our understanding of this lesser-studied type of expressions and provides novel insights into their semantic and discourse behavior. Furthermore, it will be expected to facilitate the exploration of other similar expressions from these perspectives.

1. Introduction

There are nominal expressions accompanied by the *as*-clause that includes the passive form of the verb *know*, as in (1a). There are also corresponding expressions with similar configurations except voice, as in (1b).

(1)[1] a. The Huffington Post knows its way around search engine optimization, or S.E.O. as it's known. (COCA, Newspaper, 2010)
b. IT IS A FAIRLY SURE BET THAT "WELFARE as we know it" will end. (COCA, Magazine, 1994)

Both examples in (1) have in common the connective *as*, the verb *know*, and the pronoun *it* referring to the noun immediately before the *as*-clause. Because these two expressions are similar in terms of their configurations and the lexical items used, their behavioral patterns are also expected to be similar. This paper demonstrates that although they indeed have something in common, their behavior as a whole clearly differs from the expectation. In particular, in comparison with the *NP as we know it* expressions as in (1b), this study investigates (i) whether the *NP as it is known* expressions as in (1a) have strong associations with particular situations, (ii) what semantic/functional properties they have, and (iii) whether they have any discourse properties.

This paper is organized as follows: Section 2 reviews two studies in the literature; Section 3 introduces the methodology of a corpus investigation; Section 4 provides and considers its results, focusing on three points (i.e., relevance to specific situations, sematic/functional properties, and a role in discourse); and Section 5 concludes and provides directions for further research.

2. Literature Review

This section reviews two studies concerning two kinds of expressions in (1), and attempts to show some of their basic characteristics. The first study, Sato (2023), deals with the expressions as in (1b) and the second one, Lee-Goldman (2006), describes those in (1a).

2.1. Sato (2023)

Among previous studies[2], Sato (2023) conducted the most detailed research on the expressions as in (1b) using the Corpus of Contemporary American English (COCA). Sato's (2023) focus was on 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*, *his life as he knows it*), hereafter referred to collectively as the *NP as we know it* expressions.[3] I investigated their behavior in relation to situations where their referents are involved and revealed the highly skewed distribution of the *NP as we know it* expressions by situational type.

Specifically, Sato (2023) demonstrated that the referent of a nominal modified by the *as*-clause markedly tends to participate in three situations: "non-existence," "origination," and "transformation." These situational types are defined as follows:

(2) **non-existence**: an entity is (going to be) absent or in crisis.
origination: an entity originates or is born.
transformation: an entity changes, or some aspect of it is altered.
(slightly adapted from Sato, 2023, pp. 76–78)

"An entity" in these situational types is intended to correspond to the referent of a nominal modified by the *as*-clause. And a typical example of each situational type is illustrated in (3).

(3) a. The world as we know it will <u>cease to exist</u>. (non-existence)
b. Henry Steinway <u>invented</u> the piano as we know it. (origination)
c. Television as you know it is about to <u>change</u>. (transformation)
(Sato, 2023, pp. 76–78)

The subject referent in (3a) is depicted to disappear in the future, (3b) represents the origin of the object referent, and (3c) indicates that some aspects of the subject referent will change. Therefore, these examples are categorized into "non-existence," "origination," and "transformation," respectively.

Sato (2023) classified all the examples retrieved by corpus research and found out that the above three situational types accounted for approximately 70%. Surprisingly, "non-existence" alone accounted for approximately 50%. The results are shown in Table 1.

Moreover, I compared typical variants (i.e., *NP as we know/knew it/them*) of the *NP as we know it* expressions with similar expressions with the relative pronoun *that* or *which* (i.e., *NP that/which we know/knew*) with respect to situational types; the results

Table 1. Distribution of *NP as we know it* Expressions with Respect to Situational Types

Situational type	Number of *NP as we know it* expressions
Non-existence	793 (49%)
Origination	161 (10%)
Transformation	160 (10%)
Others	501 (31%)
Total number	1615 (100%)

Note. Slightly adapted from Sato (2023, p. 83)

(see Table 2) revealed that the former, in contrast with the latter, has significant associations with the three situational types: "origination," "transformation," and especially "non-existence" ($\chi^2(3) = 391.577$, $p < .01$, $V = 0.486$).

Table 2. Distribution of Two Types of Expressions with Respect to Situational Types

	Non-existence	Origination	Transformation	Others	Row total
NP as we know/knew it/them	692	158	133	398	1381
NP that/which we know/knew	8	10	3	255	276
Column total	700	168	136	653	1657

Note. Slightly adapted from Sato (2023, p. 84)

Given these characteristics, we are prompted to ask *Does its passive counterpart (i.e., the* NP as it is known *expressions) exhibit similar behavior*? We will answer that question in Section 4.1.

2.2. Lee-Goldman (2006)

According to our literature review, the detailed research on the *NP as it is known* expressions is scant. For example, Lee-Goldman (2006) (also cf. Huddleston and Pullum, 2002) mentioned the expressions when discussing the gap within some types of *as*-clauses.[4] However, his focus was not on semantic/functional aspects of the *NP as it is known* expressions per se but on syntactic aspects of expressions accompanied by parenthetical *as*, including the *NP as it is known* expressions (and those shown in note 4).

Examples relevant to this paper are those with what he called name-*as*[5]:

(4) a. Logical addresses, or IP addresses as they are known as in the computer world, are destined to be hacked.
b. Melissa, or ChildOfBabylon [sic. Child Of Babylon] as I know her, is one of my old online journal friends.
c. They were "switched," as dealership salespeople refer to it.
d. In October, MK, as her friends call her, took a leave of absence from NYU [...].
(slightly adapted from Lee-Goldman, 2006, pp. 5-6))

Lee-Goldman (2006) pointed out that by virtue of this type of *as*-clauses, the NP that immediately precedes each of the name-*as* clauses in (4) refers to its referent's name rather than its referent per se. To put it another way, it is meta-linguistically used. For example, (4a) shows that logical addresses are known as "IP addresses" and (4b) intends to indicate that the speaker knows Melissa's another name, "Child Of Babylon." The verbs *refer* and *call* can also be used in this type of *as*-clauses, as in (4c-d). Although these two examples, unlike (4a-b), do not have overt expressions corresponding to those modified by the name-*as* clause, all these four examples convey the name of a person, a thing, or a situation. Because we are concerned with the elucidation of the specific properties of the *NP as it is known* expressions through comparison with the *NP as we know it* expressions, we focus on cases as in (4a-b).[6] Based on these two examples, our question is as follows: Although both indeed involve name-*as*, is there any semantic/functional difference between them?

In what follows, the issues of the *NP as it is known* expressions arising from Sato (2023) and Lee-Goldman (2006) are assessed based on quantitative corpus research. The data collected from the corpus will reveal specific characteristics and uses of the expressions in question.

3. Methodology of Corpus Investigation

This section introduces how I used the Corpus of Contemporary American English (COCA) to investigate the behavior of the *NP as it is known* expressions. First, I used the following search query to retrieve the data: _n* as_cs it/they/I/he/she/you/we BE known.[7] *_n** and *as_cs* are abbreviations in the COCA for any type of noun and *as* as a subordinating conjunction, respectively. The copular *be* in capital letters includes all its variants (e.g., *is*, *was*, *were*), and each oblique slanting line inserted between pronouns means *or*.

Second, this investigation was conducted to be compared with the *NP as we know it* expressions; with respect to the data of the *NP as we know it* expressions, we owe it to Sato (2023). In Sato's (2023) investigation, genres such as "TV/MOVIES," "BLOG," and "WEB-GENL" were excluded because it was conducted before the update of the COCA to expand its scale to incorporate them. Thus, to conduct a study under the same conditions, we also excluded those data.

Last, the examples targeted in this study are those in which the referent of the initial NP in the *NP as it is known* expressions corresponds to that of the pronoun within the *as*-clause that seems to modify the NP. Other cases were removed by hand as noises.[8] As a result, 94 tokens were retrieved. The next section presents the results of this investigation and discusses them in detail.

4. Results and Discussion

This section presents the results of the corpus investigation introduced above and discusses them from three perspectives: preference for specific situations, semantic/functional properties, and a role in discourse.

4.1. Association with Specific Situations

This section addresses the extent to which the *NP as it is known* expressions tend to be used in three situational types: "non-existence," "origination," and "transformation." The results of the classification are shown in Table 3, with relevant examples below it.

(5) a. […] the president could end American participation in the Trans-Pacific Partnership. its [sic. It's] fair to assume that the TPP as it's known is now <u>dead</u>. (non-existence) (COCA, Spoken, 2016)

b. Georg Friedrich Hindel, or George Frideric Handel as he is known today, was <u>born</u> in 1685 in Halle, Germany. (origination) (COCA, Magazine, 1999)

c. With Kennedy, he energized the Atlanta Opera as it's known today. (transformation) (COCA, Newspaper, 2005)

d. The Arvin Federal Camp, or Weedpatch camp [sic. Camp] as it was known, still stands at the edge of town. (others)
(COCA, Newspaper, 1999)

e. Early on the morning of Jan. 31, 1968, as Vietnamese celebrated the Lunar New Year, or Tet as it is known locally, […]. (others)
(COCA, Newspaper, 2018)

f. […] whereby landscape refers to both space and thoughts within and about "the world as it is known to those who dwell therein" (1993:156). (others)
(COCA, Academic, 2007)

Table 3. Distribution of *NP as it is known* Expressions with Respect to Situational Types

Situational type	Number of *NP as it is known* expressions
Non-existence	9 (9%)
Origination	9 (9%)
Transformation	3 (3%)
Others	77 (79%)
Total number[9]	98 (100%)

Example (5a) shows that the subject referent (i.e., the TPP) of the *NP as it is known* expression already does not exist, and (5b) indicates the date and place of birth about the subject referent (i.e., George Frideric Handel); therefore, their categories are "non-existence" and "origination," respectively. With example (5c), the verb *energize* suggests that the two persons helped make the object referent (i.e., the Atlanta Opera today) popular and thus it is possible to assume that some change in its aspects was caused before; accordingly, this example is categorized into "transformation."[10] The category "others" comprises the other situational and non-situational types. Example (5d) conveys that the subject referent (i.e., Weedpatch Camp) *exists* in some area, and the object referent (i.e., Tet) in (5e) is the entity of being *celebrated*; accordingly, these examples do not express any of the three situational types (i.e., "non-existence," "origination," and "transformation") and thus are classified into the "others" category. Various verbs (e.g., *convict*, *teach*, *include*, *commit*) are used with the expressions, and no striking commonalities were observed regarding the types of situations within this category. The *NP as it is known* expression in (5f) functions as the object of the prepositions *within* and *about*, and its referent per se does not directly participate in the situation described by the matrix verb *refer*. It functions as a place (e.g., *the house/room* as in *within the house/room*) where space and thoughts exist and serves as a particular subject (e.g., *philosophy/culture* as in *a book about philosophy/culture*); thus, the category of this example is "others" as well.

Importantly, among all these types of situations, "non-existence" and "origination" each account for 9% and "transformation" for 3%. Together, they make up approximately only 20% of the total. Remember the distribution of the *NP as we know it* expressions, where "non-existence" accounts for approximately 50% and "origination" and "transformation" each for 10%, together making up 70 % of the total. What is particularly striking is that the rate of "non-existence" in the *NP as it is known* expressions is less than one fifth of the *NP as we know it* expressions. Considering their similar configurations and common lexical items, this difference between these two kinds of expressions is surprising. To further strengthen our results, we compared Tables 1 and 3 by conducting an χ^2 test.[11] The results revealed significant differences among conditions ($\chi^2(3) = 97.966$, $p < .01$, $V = .239$). Interestingly, the residual analysis after the χ^2 test showed that the *NP as it is known* expressions significantly tend to avoid "non-existence" whereas the *NP as we know it* expressions significantly prefer it at the $p < .01$ level. These results are represented in Table 4.

Table 4. Distribution of Two Types of Expressions with Respect to Situational Types

	Non-existence	Origination	Transformation	Others	Row Total
NP as we know it expressions	793 ($p < .01$)	161	160	501 ($p < .01$)	1615
NP as it is known expressions	9 ($p < .01$)	9	3	77 ($p < .01$)	98
Column Total	802	170	163	578	1713

Given the category "others" is a miscellany of various situations and non-situations, these results demonstrate that the *NP as it is known* expressions are highly unlikely to have a strong association with particular types of situations (especially with "non-existence"), unlike the *NP as we know it* expressions. This implies that there must be some other common characteristic aspects within the *NP as it is known* expressions. In the subsequent sections, therefore, we aim to identify and present such properties.

4.2. Naming Use

In the corpus investigation, we identified three types of properties of the *NP as it is known* expressions. The first property concerns what Lee-Goldman (2006) called name-*as*. As we saw in Section 2.2, the name-*as* clause serves to indicate that the immediately preceding NP is meta-linguistically used. In (6a), for example, the *as*-clause shows that the word "TPP" is pronounced (or written) as three individual letters—T-P-P. Furthermore, within the total number of the *NP as it is known* expressions retrieved, 66% of them (i.e., 62/94 tokens) both involve name-*as* and emerge in specific configurations exemplified in (6).

(6) a. […] the president could end American participation in the Trans-Pacific Partnership. its [sic. It's] fair to assume that the **TPP** as it's known is now dead. (= (5a))

b. The University of California, Santa Barbara—**UCSB** as it's known—could hardly be more different or farther away from Concord, Mass., where […]. (COCA, Newspaper, 1990)

c. There's rookie Jason Williams, or **White Chocolate** as he's known now, a playground version of Pete Maravich. (COCA, Newspaper, 1999)

d. We also had a dolphin, or **dorado** as it's known in Spanish, in the box. (COCA, Magazine, 2007)

e. The underground church, **the house churches** as they're known, have been under renewed persecution recently. (COCA, Spoken, 2000)

Typically, an expression (in bold) with name-*as* is found after the corresponding expression (underlined) in the immediately preceding sentence (e.g., (6a)) or in the same sentence (e.g., (6b-e)). For the latter, there are some variations: with an em-dash, (e.g., (6b)), a comma and *or* (e.g., (6c-d)), and only a comma (e.g., (6e)). Notably, all 62 examples are realized in one of these two distributional patterns.

Given these distributional properties, the main and typical property of the *as*-clause of this type is to signal that an expression accompanied by this *as*-clause is a paraphrase of another one; accordingly, "TPP" and "UCSB" are initialisms for "Trans-Pacific Partnership" and "(The) University of California, Santa Barbara" in (6a-b); "Jason William" in (6c) is paraphrased with his nickname "White Chocolate"; "dorado" in (6d) is a Spanish name for a kind of surface-dwelling fish called “dolphin”; and “The underground church” in (6e) is paraphrased with another name “the house churches.” As can be seen from these examples, various kinds of paraphrasing are available with the use of name-*as*.[12]

Of course, name-*as* can be used to name the preceding NP, irrespective of the presence of any paraphrased expression; however, only few examples (i.e., 7 tokens) illustrate this point, some of which are exemplified in (7).

(7) a. He’ll go to the truce village as it’s known along the border and what follows? (COCA, Spoken, 2019)
b. This place is separate both from the Own [sic. Orun] inhabited by the Orixs [sic. Orixas] and Bahia as it is known to Bahians. (COCA, Academic, 2003)
c. As it does it creates clouds of course. Then you have pyrocumulus clouds as they’re known across the streets. (COCA, Magazine, 2007)

There is no expression comparable to “the truce village,” “Bahia,” or “pyrocumulus clouds” within the given and immediately preceding sentences (and actually, up to three sentences earlier) in each text. The word “clouds” per se appears in the first sentence in (7c), but pyrocumulus clouds are a subtype of clouds; thus, the two are not the same. The other examples (i.e., 25 tokens) are not of name-*as* or at least ambiguous between this and another type of *as* (i.e., the restrictive use of *as*). The latter use functions to qualify and constrain what is denoted by the noun-phrase preceding restrictive-*as*. A typical example is the phrase “the world as I know it,” in which the *as*-clause narrows down the scope of the preceding entity (i.e., the world). We will discuss this use in the next section.

From these results, the number and rate (i.e., 69/94 tokens and 73%, respectively) of clear examples of name-*as* in the *NP as it is known* expressions appears to be very high. It is still unclear, however, whether this is a unique property of the *NP as it is known* expressions. It is possible that the same proportion is also observed in similar expressions, the *NP as we know it* expressions. To exclude such a possibility, we conducted the same kind of investigation for the *NP we know it* expressions as well. In some cases, it is potentially difficult to determine definitely whether *as* in these expressions is of name-*as*; it may also be of the restrictive use. To deal with this issue with maximal accuracy, we examined the degree to which paraphrased expressions co-occur with them, because, as shown above, the presence of paraphrased expressions is the most typical environment in which name-*as* appears. To ensure comparability, we conducted our investigation under the same conditions as that of the *NP as it is known* expressions: whether there is an overt paraphrased expression before its paraphrasing expression within the same sentence or the immediately preceding sentence. Our results showed that most of the examples were of the restrictive use and the number of the naming use (even including the potentially relevant ones) was only 49 (of 1580) tokens, accounting for only 3%. Some of them are given in (8).

(8) a. Well, MSF, or **Doctors Without Borders** as we know them here in the United States, has been in Afghanistan providing medical care during several wars, […]. (COCA, Spoken, 2004)

b. He issued another executive order yesterday, this one for the Food and Drug Administration, the **FDA** as we know it, to examine ways to shore up prescription drug shortages. (COCA, Spoken, 2011)

c. Psychologically Laurie, **the child** as you knew her, withdrew from the pain and fear […]. (COCA, Fiction, 1992)

d. […] so many of the great political upheavals […] were launched from the territory west of the Ohio River. **The region** as they knew it was what gave the country Socialists […]. (COCA, Magazine, 2004)

Examples (8a-b) seem to correspond to examples (6a-b); “MSF” in (8a) is an initialism for the French phrase ‘Médecins Sans Frontières,’ which is translated into ‘Doctors Without Borders’ in English; “FDA” in (8b) is an initialism for “Food and Drug Administration.” They pertain to how “MSF” and “FDA” may be referred to differently, and thus are regarded as examples of name-*as*.[13] Examples (8c-d) somewhat differ from them, because “Laurie” and “the territory west of the Ohio River” are highly unlikely to be referred to literally as “(the) child” and “(the) region” as their names. They seem to be cases of the restrictive use of *as*. Due to the presence of the paraphrased expressions, nevertheless, our investigation had to include these kinds of examples. Considering this point and the still remarkably low frequency of paraphrased expressions, we can safely say that the naming use of the *as*-clause in the *NP as we know it* expressions is severely limited. As a result, it is concluded that name-*as* and a paraphrasing function are the main properties of the *NP as it is known* expressions.

4.3. Restrictive Use

Next, we discuss the remaining 25 examples. They can be interpreted differently from those of name-*as*.[14] *As*-clauses in these expressions serve to restrict the range of the preceding NPs modified by them: a restrictive use. This property is the second (less typical) property of the *as*-clauses in the *NP as it is known* expressions. The following examples illustrate this point evidently:

(9) a. A ruler who doesn't want to control the political system but to break the system as it is known? (COCA, Magazine, 2019)

b. This refers specifically to the Holy Land as it is known within Christian tradition, […] (COCA, Academic, 2018)

c. […] whereby landscape refers to both space and thoughts within and about "the world as it is known to those who dwell therein" (1993: 156). (= (5f))

They are clearly not metalinguistic uses of the NPs preceding the *as*-clauses. "The system" refers to "the political system" in (9a), but is not generally called or recognized as "(the) system" literally. "The system as it is known," as a whole, carries a meaning similar to "the existing system." Similarly, it is more likely that the *as*-clause in (9b) does not prompt the metalinguistic interpretation of "(the) Holy Land" but specifies one version or type of Holy Land (e.g., how the Holy Land is known by Christians as opposed to Buddhists). The same can be said of (9c), where a specific realm of perception of the world by a certain group of people (i.e., those who dwell therein) is intended rather than the entire world. Therefore, *as*-clauses in all these examples modify the preceding NPs and narrow their scope.

Moreover, this type of *as*-clause and the preceding noun typically constitute one unit. A fragment answer test, a traditional constituency test, indicates that "the Holy Land as it is known within Christian tradition" serves as a (phrasal) constituent[15]:

(10) A: This refers to the Holy Land as it is known within Christian tradition.
B: Sorry, what did you say? What Holy Land?
A: The Holy Land as it is known within Christian tradition.

In a response to the question from speaker B, speaker A can use the entire phrase "the Holy Land as it is known within Christian tradition" as an answer. Given that this question aims to elicit a kind of Holy Land, the *as*-clause here can be interpreted as serving to restrict the scope of the notion. In other words, it specifies some property/aspect of the Holy Land in question (cf. Huddleston and Pullum, 2002, p. 1150). Another piece of evidence for this is that double quotation marks are used in (9c). Notice that they encompass not only the "the world," but also the entirety of "the world as it is known to those who dwell therein." This implies that the writer of this sentence does not consider "the world" and "as it is known to those who dwell therein" as separate and discrete, but (s)he considers them as a single constituent.

Notably, these two characteristics are also found in the *NP as we know it* expressions. Let us first consider (11).

(11) A: Then, the world as he knew it was over.
B: Sorry, what did you say? What world?
A: The world as he knew it.

An interlocutor (i.e., speaker B) can inquire, "What world?" when hearing another person (i.e., speaker A) say, "The world as he knew it was over." And in reply, speaker A can answer with "The world as he knew it," which refers to a specific part of the whole world. This kind of answer is typically possible for the *NP as we know it* expressions, because most of them are of the restrictive use. Furthermore, some examples with double quotation marks were found in the expressions, as in (1b), repeated as (12).

(12) IT IS A FAIRLY SURE BET THAT "WELFARE as we know it" will end.
(= (1b))

Once again, the entire phrase "WELFARE as we know it" is encompassed by the double quotation marks.

However, examples demonstrating paraphrases, as in (6), do not seem to exhibit such characteristics. Regarding (6a), for example, if an interlocutor inquired “What TPP?” in the context where the sentences in (6a) are uttered, another interlocutor would feel confused and never say “the TPP as it’s known.” Consequently, the following question and its answer are pragmatically very unnatural.

(13) A: The president could end American participation in the Trans-Pacific Partnership. It’s fair to assume that the TPP as it’s known is now dead.
B: Sorry, what did you say? What TPP?
A: The TPP as it’s known.

Interestingly, no examples with name-*as* were found with double quotation marks placed at the ends of the string *NP* as-*clause*.

To sum up, there are two uses in the *NP as it is known* expressions: naming and restrictive uses. Quantitative analysis revealed that the former use is far more prevalent than the latter use (i.e., 69 vs. 25 tokens). Conversely, the *NP as we know it* expressions exhibited the opposite distribution: the naming use is scarce while the restrictive use is dominant (i.e., 49 vs. 1531 tokens). Table 5 shows these differences between the two types of expressions in question.

Table 5. Prevalence of Two Uses between Two Types of Expressions

	Naming use	Restrictive use
NP as it is known expressions	Main	Limited
NP as we know it expressions	Limited	Main

We can see form this table that these types of expressions exhibit a mirror relation to each other and furthermore, the distinct distributional pattern found within each type of expressions constitutes its own unique characteristics.[16]

4.4. A Role in Discourse

The third property is that *NP* in the *NP as it is known* expressions, especially when they involve name-*as*, tends to be overtly realized again and to be a primary topic. This tendency is evident in (14).

(14) Darlene had lasted only a few months before she'd been replaced by Callie Kreutzer, an art student at the Dorset Academy, who happened to be the girlfriend of Justy Junior—or June as he was known. **June**, age twenty-four, worked for his dad as a salesman. (COCA, Fiction, 2011)

Initially, June, an alias for Justy Junior, is introduced as the object referent of the preposition *of*. He then serves as the subject referent in the following sentence, where "June" occupies the topic position and the verbal phrase the comment position (cf. Lambrecht, 1994). This section explores the correlation between topicality and the *NP as it is known* expressions.

To examine such a discourse factor, the current study utilized the following methodology. Since the name-*as* clause in these expressions mainly serves to signal that the NP accompanied by this *as*-clause is a paraphrase of another one (see Section 4.2), a total of 62 examples, such as (6), demonstrating this paraphrastic relationship were selected for analysis. With these examples, an analysis was conducted to ascertain the specific grammatical case through which the NP in question was reintroduced, as case marking is closely related to topicality. Givón, for example, provides the following topicality order: SUBJECT > DIRECT OBJECT > OTHERS (1990, p. 901). Given that the subjects and objects are commonly linked with nominative and accusative cases in English, this paper posits that the two grammatical cases typically represent a primary topic, classified as "nominative/accusative." The category of other cases (e.g., possessive, dative) and part of a phrase (e.g., *welfare* in *welfare reform*) is "others," and if the NP in question is not realized again within a specified range, it is categorized into "ϕ."[17]

The specific procedure, mainly based on Givón (1983), was as follows: The basic unit to count was fixed as a clause; the subordinate clause was considered part of the matrix clause[18] and the coordinate clause was counted as an independent clause. The scope of the search was experientially limited to a maximum of three clauses subsequent to the clause including the *NP as it is known* expressions; the counting persisted even in the event of a change in the speaker or paragraph.[19] This decision was made due to the likelihood that the topicality of the NP in question may persist or that the NP may acquire such a property anew in utterances by another interlocutor or within a new paragraph. Regarding grammatical cases, only the first case in which the NP reappeared was counted; thus, it was counted once, even if it reappeared multiple times within the following three clauses. The same procedure and analysis were conducted for the *NP as we know it* expressions for comparison. The number retrieved was the same as that of the *NP as it is known* expressions and was extracted from 1580 tokens using stratified sampling.[20] After categorizing them into "nominative/accusative," "others," and "ϕ," these two types of expressions were compared by conducting an χ^2 test.

The results are presented in Table 6, with relevant examples of the two types of expressions below it:

Table 6. Distribution of Two Type of Expressions with Respect to Grammatical Cases

	Nominative/Accusative	Others	ϕ	Row Total
NP as it is known expressions	23 ($p < .01$)	25	14 ($p < .01$)	62
NP as we know it expressions	7 ($p < .01$)	14	41 ($p < .01$)	62
Column Total	30	39	55	124

NP as it is known expressions:

(15) a. The first Europeans found the Australian landscape, or bush as it was known, to be unfamiliar, hostile, and lonely. **It** was an alien place in which to live. (nominative) (COCA, Academic, 1990)

b. Paul, or Saul as he was known before converting to Christianity, is reported to have had a fit that resembled an epileptic seizure: "a light from the sky suddenly flashed around **him**. He fell […]. (dative: others) (COCA, Academic, 2013)

NP as we know it expressions:

(16) a. Later this year, we will offer a plan to end welfare as we know it… We have to end **welfare** as a way of life and make it a path to independence and dignity… (accusative) (COCA, Magazine, 2004)

b. […] the fairy tales I read as a girl have no relationship to marriage as I know it—my own parents [sic. parents'] marriage has little relationship to **marriage as I know it**. (dative: others) (COCA, Magazine, 1996)

With the *NP as it is known* expressions, a landscape called "the Austrian landscape" or "bush" is realized again as the subject of the second sentence in (15a), and a person called "Paul" or "Saul" is realized again as the object of the preposition *around* in

(15b). With the *NP as we know it* expressions, “welfare (as we know it)”[21] in (16a) is realized again as the verbal object in the following sentence, and in (16b) “marriage as I know it” is realized again as the object of the preposition *to*.

The results (Table 6) revealed significant differences among conditions ($\chi^2(2) = 24.890$, $p < .01$, $V = .448$). Moreover, the residual analysis after the χ^2 test suggested that *NP* in the *NP as it is known* expressions prefers to be realized again in any manner, as indicated by the significantly lower frequency for the “ϕ” category in these expressions (at the $p < .01$ level). This implies that saliency or focus of attention (cf. Langacker, 2008) persists in this NP. More specifically, as evidenced by the significantly higher frequency for the “nominative/accusative” category in the expressions (at the $p < .01$ level), *NP* in the *NP as it is known* expressions, as opposed to the *NP as we know it* expressions, tends to represent a primary topic. From these results, it is concluded that the use of the name-*as* clause in the *NP as it is known* expressions signals that the modified NP tends to be reintroduced as a primary topic in the following discourse.

5. Conclusion and Outlook

This study revealed several key distributional properties of the *NP as it is known* expressions, as summarized in Table 7.

Table 7. Summary of Distributional Properties of Two Types of Expressions

	Preference for particular situations (e.g., “non-existence”)	Naming use	Restrictive use	Primary topic
NP as it is known expressions	Limited	Main	Limited	Main
NP as we know it expressions	Main	Limited	Main	Limited

As can be seen from the table, the *NP as it is known* expressions, unlike the *NP as we know it* expressions, have no significant connections with particular situations, such as “non-existence.” The *as*-clause in the former expressions is instead mainly used to name the preceding NP and typically signals that this NP is a paraphrased expression. Moreover, the *as*-clause in question can also restrict the range of the preceding NP and in that case, each of the *NP as it is known* expressions typically constitutes a unit, or phrasal constituent. With a discourse aspect, *NP* in the *NP as it is* *known* expressions significantly tends to be overtly realized again in the subsequent sentences and is likely to be a primary topic in the following discourse. Interestingly, in all these respects, these two types of expressions exhibit a mirror relation to each other.

These results will be able to facilitate the comparison of similar expressions, such as *NP known as* expressions and *NP as we call/put/refer to it* expressions, to reveal their (di)similarity. From a wider discourse perspective, topic persistence or a topic-promoting function (cf. Lambrecht, 1994) may also be a crucial distinguishing factor. Consequently, further scrutiny is warranted to unveil their distinct properties. In this regard, this study provides the new insights into the exploration of the *NP as it is known* expressions and other several similar expressions, contributing to future linguistic inquiry.

Acknowledgements

I am grateful to 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. Emphasis is mine and the same applies hereafter.
2. See Ishibashi (1966), Watanabe et al. (1976, 1981, 1995), Kanaguchi (1978), Kinugasa (1979), Ogawa (1985), Hirota (1988), Kumagai (1989) and Yagi (1996). These previous studies do not restrict their focus to expressions where the *as*-clause includes the verb *know*, but they consider the general description of a kind of *as* that modifies the preceding nominal.
3. The reason for using this name is that the variant *NP as we know it* is the most typical and frequent form among all the variants; it accounts for over 75% of the total (Sato, 2023, p. 75).
4. Other examples illustrating this point are as follows:
   (i) As is often the case ___, most people don't understand the important issues.
   (ii) As most people do ___, she probably initially thought that it would be easy.
   (Lee-Goldman, 2006, p. 4)
   The gap in the former sentence corresponds to the matrix clause in (i) as a whole and the one in the latter to the verbal phrase in the matrix clause in (ii).
5. One characteristic of name-*as* is its syntactic mobility:
   (i) The best thing is my ability to "vague out", *as I call it*.

(ii) This is the account I attribute to, *as I call him*, 'Satre-Two'.

(iii)* Trish has been working too hard, *as I call her for short*.

(Lee-Goldman, 2008, p. 243)

In (i), the name-*as* clause appears after a phrase it modifies and it can also appear before it, as in (ii); however, it cannot appear anywhere else, according to Lee-Goldman (2008). This is an interesting syntactic behavior and is also interesting in terms of grammaticalization, but since this paper tries to seek out the semantic/functional properties of the more specific expressions (i.e., the *NP as it is known* expressions), this syntactic mobility is beyond the scope of it and not treated here. For interested readers, see also Lee-Goldman (2007, 2012).

6. To avoid complicating the discussion and the potential influence that might be caused by the presence of commas (cf. Yagi, 1996), we only treat comma-less variants and exclude examples as in (i) and (ii).

(i) "Live TV, as we know it, is over," she says. (COCA, Magazine, 2004)

(ii) CBD or cannabidiol, as it's known scientifically is one of dozens of compounds in marijuana called cannabinoids. (COCA, Newspaper, 2014)

Assessing the difference and similarity in their behavior is left to my future research.

7. First access to the data was June 6, 2021.

8. For example, the following examples are excluded that express the comparison between the matrix clause and the *as*-clause:

(i) As respected for her candor as she's known for her humor, Goldberg once set the record straight […]. (COCA, Magazine, 2001)

(ii) […], all of them known to Jeannette as she was known to them, […].

(COCA, Fiction, 1995)

9. The total number in this table is more than that of all the tokens retrieved, because the referent of the head of the *NP as it is known* expressions can participate in more than one situation when the two (or more) predicates or prepositions are conjoined within the matrix clause in a sentence:

(i) The first Europeans found the Australian landscape, or bush as it was known, to be unfamiliar, hostile, and lonely. (COCA, Academic, 1990)

Here, the three predicates *unfamiliar*, *hostile*, and *lonely* are used to describe the properties that the Australian landscape (known as bush) has; thus, each of them is categorized into the "others" category.

10. This is less typical than (3c), which is a typical example of "transformation" in that the change is directly expressed linguistically (i.e., by the verb *change*). Another example, shown in below, is also less typical, because the change in physical state is temporary and expected to return to normal soon.

(i) In the late 1880s, Wovoka, or Jack Wilson as he was known in the non-Indian world, fell ill with a severe fever, which happened to coincide with a solar eclipse.

(COCA, 2016, Academic)

In fact, there were no typical examples of "transformation" in the *NP as it is known* expressions.

11. Data analysis was performed using data analysis software, js-STAR, available online at https://www.kisnet.or.jp/nappa/software/star/.
12. An anonymous reviewer points out the following variant, where an adverb is inserted between *is* and *known*, as opposed to e.g., (5e), and wonders whether the *NP as it is known* expressions are now becoming fixed phrases.

    (i) High-Performance Computing, or HPC as it is <u>generally</u> known, has been a mainstay of infrastructure and hardware engineering groups for decades.

    (https://www.rxdatascience.com/high-performance-computing)

    Without examining historical changes, however, it is difficult to make a final judgment, but currently, it seems that such fixation has not occurred and adverbs can arise relatively freely, which is supported by the fact that over 30 examples of this kind are found in COCA.
13. Two informants also judged these two examples to be of name-*as*.
14. All these examples were provided to two informants with a sufficient context of approximately 100 words, including the *NP as it is known* expressions. Then, both informants judged that the restrictive use is preferrable as an interpretation of these expressions, although the naming use may also be possible or ambiguous in some of them.
15. This test is not effective against some examples of the restrictive use, not because they are not constituents, but because they do not have enough specificity. "The system as it is known" in (9a), for example, was not judged to be a likely answer to the question "What system?" Instead, "The political system as it is known" can qualify as the answer, which also suggests that the string *NP* as-*clause* forms a constituent.
16. An anonymous reviewer poses the question of why the passive form is deeply related to name-*as*. It seems that it often does not matter specifically who knows/refers to an entity by a particular name, but rather how it is generally or conventionally named. In fact, all of the *NP as it is known* expressions with name-*as* do not explicitly involve a specific individual who knows an entity's name (e.g., *NP as it is known by John*). Instead, in one case (e.g., (6a-b)), it is assumed that many people know its name, rendering it unnecessary to specify who knows it, while in another case (e.g., (5e-f)), it is either implied or indicated obliquely that a group of people know it. These situations mesh well with the well-known function of passive voice: defocusing agent (cf. Shibatani 1985). Therefore, the *NP as it is known* expressions, unlike the *NP as we know it* expressions, are likely to be used with name-*as*.
17. A similar methodology and perspective can be observed in Tokunaga (2019), who revealed that the subjects of *as*-clauses with subject-auxiliary inversion are more likely to function as primary topics in the subsequent sentences than those without such inversion.
18. This means that when the NP in question is reintroduced as the subject or object in the subordinate clause, it will be categorized as "others," not "nominative/accusative."

19. Even after excluding such cases, the results remained statistically significant.
20. All of the extracted examples showed no clear sign of name-*as* and seem to be of restrictive-*as*; thus, we should be aware that the comparison is done between the typical use/behavior of the *NP as it is known* expressions and that of the *NP as we know it* expressions.
21. Often ambiguous and hard to distinguish (especially when the subject of the *as*-clause is *we*) is which is paraphrased, *NP as we know it* or *NP*; thus, we did not distinguish between the two in this study.

References

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

Givón, T. (1983). Topic Continuity in Discourse: An Introduction. In T. Givón (Ed.), *Topic Continuity in Discourse: A Quantitative Cross-Language Study* (pp. 1–41). John Benjamins. https://doi.org/10.1075/tsl.3.01giv

Givón, T. (1990). *Syntax: A Functional-Typological Introduction (Vol II)*. John Benjamins. https://doi.org/10.1075/z.50

Hirota, N. (1988). *As*-setsu to Kankeishi. *Eigo Seinen (The Rising Generation)*, *133* (11), 9. https://doi.org/10.11501/4435575

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press. https://doi.org/10.1017/9781316423530

Ishibashi, K., Hirose, T., Ito, K., Takanashi, K., Tori, T., & Watanabe, T. (Eds.). (1966). *A Dictionary of Current English Usage I*. Taishukan. https://doi.org/10.11501/8312728

Kanaguchi, Y. (1978). *Gendai Eigo no Hyougen to Gokan*. Taishukan. https://doi.org/10.11501/12580610

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

Kumagai, Y. (1989). A Note on English *As*-clause. *Mita Working Papers in Psycholinguistics 2*, 55–63.

Lambrecht, K. (1994). *Information Structure and Sentence Form*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620607

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

Lee-Goldman, R. (2006, September 29). *Parenthetical* As *(and Movement Paradoxes)*. Syntax & Semantics Circle, Berkely, California, United States.

Lee-Goldman, R. (2007, March 8-11). *The Relative Proform* As. Georgetown University Roundtable on Language and Linguistics, Georgetown, Washington DC, United States.

Lee-Goldman, R. (2008). Supplemental Relative Clauses and Syntactic Generality. *BLS*, *34*, 233–244. https://doi.org/10.3765/bls.v34i1.3572

Lee-Goldman, R. (2012). Supplemental Relative Clauses: Internal and External Syntax. *Journal of Linguistics*, *48* (3), 573-608. https://doi.org/10.1017/S0022226712000047

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

Sato, Y. (2023). On the *NP as we know it* Expressions and its Variants. *English Corpus Studies*, *30*, 71–94.

Shibatani, M. (1985). Passive and Related Constructions: A Prototype Analysis. *Language*, *61*, 821–848. https://doi.org/10.2307/414491

Tokunaga, K. (2019). Zuiiteki na Shugo Jyodoushi Touchi ga Tekiyou Sareta *As*-setsu no Bunmyaku teki Kinou to Sono Tokucho—Touchi Gensho to Bunmyaku teki Kinou no Interface. *Studies in Pragmatics*, *21*, 139–160.

Watanabe, T., Ando, S., Fukumura, T., Kawakami, M., Konishi, T., Miura, S., & Soranishi, T. (Eds.). (1976). *A Dictionary of Current English Usage II*. Taishukan. https://doi.org/10.11501/12580592

Watanabe, T., Ando, S., Fukumura, T., Kawakami, M., Konishi, T., Miura, S., & Murata, Y. (Eds.). (1981). *A Dictionary of Current English Usage III*. Taishukan. https://doi.org/10.11501/12580590

Watanabe, T., Fukumura, T., Kawakami, M., Konishi, T., & Murata, Y. (Eds.). (1995). *A Dictionary of Current English Usage IV*. Taishukan. https://doi.org/10.11501/12580591

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

（佐藤　嘉晃　京都大学（非常勤講師））

「ソフトウェア紹介」

音声・映像コーパス構築ツール
「Speech Indexer」の紹介

後藤　一章

1. はじめに

一般に，通常の文書コーパスと比較し，音声・映像コーパスの構築には多大なコストがかかり，未だ広く実践されているとは言い難い。ICNALE-Spoken（Ishikawa, 2023）や TCSE（Hasebe, 2015）等によってその有用性や可能性は認識されつつあるものの，音声の文字起こし作業やその後の検索方法が課題となり，個々の研究者が音声・映像コーパスを自ら構築することは容易ではなかった。

本研究は，こうした問題の解消を目指し，音声・映像コーパスの構築を支援するツール「Speech Indexer」を開発した。Speech Indexer は OpenAI の「Whisper」と呼ばれる音声認識システムを利用し，音声の自動文字起こしと検索文字列の該当箇所再生機能を有する Windows プログラムである。

本稿では，まず Whisper の概要と，その派生プログラムである「Whisper.cpp」について紹介し，続けて Speech Indexer の機能と操作方法について簡潔に述べる。さらに，Speech Indexer を使用し，英語学習者が話す英語の音声認識精度について小規模な調査を行ったため，合わせて報告する。

2. 音声認識システム

2.1 Whisper

Whisper は，OpenAI が 2022 年 9 月，MIT ライセンスのオープンソース・ソフトウェア（ソースコードを自由に利用，改変，再配布等が可能なライセンス形式）として公開した音声認識システムである。公開されて間もなくその認識精度の高さが注目を集め，オープンソースということもあり，国際的に広く使用されることとなった。また，2023 年 3 月には，有料ではあるが WebAPI と

しても公開され，同社の文章生成 AI である ChatGPT との組み合わせが容易になるなど，より柔軟にアプリケーションや Web サービスに Whisper を統合することが可能となった。

OpenAI の公式サイトによると，Whisper は Web から収集した 68 万時間分の多言語音声データを，Transformer と呼ばれる深層学習モデルによって学習している。詳細は Radford et al.（2022）に譲るが，入力された音声データはログメルスペクトログラム（log-Mel spectrogram）という特徴量に変換され，対応する文字列データ等と共に学習が行われる。認識精度は，英語で約 95.5%，日本語で約 93.6% とされている（Radford et al., 2022）。

2.2 Whisper.cpp

Whisper.cpp は，オープンソース版の Whisper を Georgi Gerganov 氏が C 及び，C++ というプログラミング言語によって新たに書き起こしたオープンソース・ソフトウェアである。オリジナルの Whisper は Python で作成されており，GPU（Graphics Processing Unit）の使用を前提としているが，Whisper.cpp では必ずしも GPU は必須ではない。プログラム全体が C/C++ で書かれていることもあり，CPU（Central Processing Unit）のみでも高速な文字起こしを実現している。ソースコードが公開されているため，コンパイル環境さえあれば OS を問わず実行が可能であり，また，FFmpeg 等の外部プログラムに依存しない点も利点として挙げられる。

2.3 Speech Indexer 開発の背景

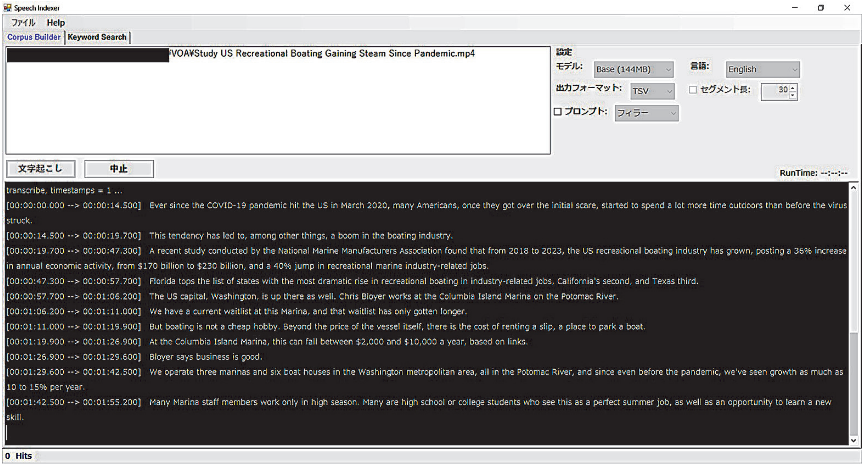
Whisper や Whisper.cpp は極めて有用なソフトウェアであるが，その利用にはコマンドラインでの操作が必須となる。コマンドライン操作は柔軟で自由度が高いが，場合によっては煩雑となる。また，ファイルの一括処理や，文字起こしテキストからの音声・映像検索等にはある程度のプログラミング知識が求められ，必ずしもコーパス言語学や外国語教育研究で手軽に活用できるとは言い難い。そこで，本研究では，基本的な認識処理はもちろん，複数ファイルの処理や，検索処理を直観的に行える GUI（Graphical User Interface）を備えたユーザフレンドリーなツールを開発することとした。

3. Speech Indexer

3.1 文字起こし

図 1 は，Speech Indexer の文字起こし画面である。操作方法は，任意の音声または映像ファイルを選択し，モデルファイル等を設定したうえで，文字起こしを実行するのみである。

音声ファイルは WAV，MP3，M4A フォーマットに対応し，映像ファイルは MP4, MOV フォーマットに対応している。Whisper.cpp は 16 ビットの WAV ファイルにのみ対応しているが，本ツールではその他の形式でも内部処理によって適切なフォーマットに変換している。



[redacted]¥VOA¥Study US Recreational Boating Gaining Steam Since Pandemic.mp4

設定
モデル: Base (144MB)
言語: English
出力フォーマット: TSV
□ セグメント長: 30
□ プロンプト: フィラー

```
transcribe, timestamps = 1 ...
[00:00:00.000 --> 00:00:14.500]   Ever since the COVID-19 pandemic hit the US in March 2020, many Americans, once they got over the initial scare, started to spend a lot more time outdoors than before the virus
struck.
[00:00:14.500 --> 00:00:19.700]   This tendency has led to, among other things, a boom in the boating industry.
[00:00:19.700 --> 00:00:47.300]   A recent study conducted by the National Marine Manufacturers Association found that from 2018 to 2023, the US recreational boating industry has grown, posting a 36% increase
in annual economic activity, from $170 billion to $230 billion, and a 40% jump in recreational marine industry-related jobs.
[00:00:47.300 --> 00:00:57.700]   Florida tops the list of states with the most dramatic rise in recreational boating in industry-related jobs, California's second, and Texas third.
[00:00:57.700 --> 00:01:06.200]   The US capital, Washington, is up there as well. Chris Bloyer works at the Columbia Island Marina on the Potomac River.
[00:01:06.200 --> 00:01:11.000]   We have a current waitlist at this Marina, and that waitlist has only gotten longer.
[00:01:11.000 --> 00:01:19.900]   But boating is not a cheap hobby. Beyond the price of the vessel itself, there is the cost of renting a slip, a place to park a boat.
[00:01:19.900 --> 00:01:26.900]   At the Columbia Island Marina, this can fall between $2,000 and $10,000 a year, based on links.
[00:01:26.900 --> 00:01:29.600]   Bloyer says business is good.
[00:01:29.600 --> 00:01:42.500]   We operate three marinas and six boat houses in the Washington metropolitan area, all in the Potomac River, and since even before the pandemic, we've seen growth as much as
10 to 15% per year.
[00:01:42.500 --> 00:01:55.200]   Many Marina staff members work only in high season. Many are high school or college students who see this as a perfect summer job, as well as an opportunity to learn a new
skill.
```

図 1. Speech Indexer の文字起こし画面

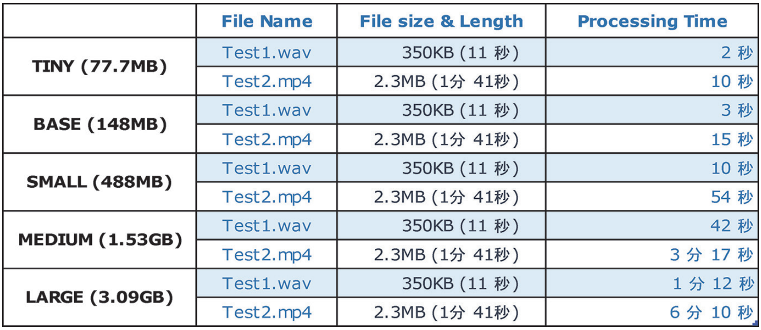
3.1.1 モデル

使用するモデルは，文字起こしの処理速度と認識精度に直接的に影響する。モデルファイルのサイズが大きければそれだけ精度は向上するが，必然的に処理時間も増加する。モデルファイルは，TINY，BASE，SMALL，MEDIUM，LARGE の 5 種類に大別される。

モデルファイルのファイルサイズと，処理に要する時間の例を表 1 に示す。データには，約 11 秒と約 1 分 41 秒の英語音声ファイルを使用した。計測は，Intel Core i5-10210U，コア数 / スレッド数：4/8，動作周波数 : 1.60GHz，RAM:

16.0 GB，の環境で行った。なお，GPU は不使用である。最小の TINY モデルでは，オリジナル音声の十分の一程度の処理時間となったが，最大の Large モデルでは，オリジナル音声の時間の 3〜6 倍程度必要となった。

表 1. モデル別による文字起こしに要する処理時間



	File Name	File size & Length	Processing Time
TINY (77.7MB)	Test1.wav	350KB (11 秒)	2 秒
	Test2.mp4	2.3MB (1分 41秒)	10 秒
BASE (148MB)	Test1.wav	350KB (11 秒)	3 秒
	Test2.mp4	2.3MB (1分 41秒)	15 秒
SMALL (488MB)	Test1.wav	350KB (11 秒)	10 秒
	Test2.mp4	2.3MB (1分 41秒)	54 秒
MEDIUM (1.53GB)	Test1.wav	350KB (11 秒)	42 秒
	Test2.mp4	2.3MB (1分 41秒)	3 分 17 秒
LARGE (3.09GB)	Test1.wav	350KB (11 秒)	1 分 12 秒
	Test2.mp4	2.3MB (1分 41秒)	6 分 10 秒

Speech Indexer には予め BASE モデルを同梱しており，ダウンロード後即座に使用できる状態となっている。ただし，高精度での認識には MEDIUM や LARGE モデルの導入が推奨され，特に日本語認識には，最低でも SMALL 以上のモデルが望ましい。

3.1.2 言語

認識言語を「English」,「Japanese」,「Auto」から選択する。「Auto」は認識速度がやや低下するが, 言語ごとに設定を切り替える作業が不要となる。また，Whisper は実際には 98 言語の音声認識に対応しており，英語と日本語以外の文字起こしも可能である。ただし，Speech Indexer の音声・映像検索機能は，英語と日本語以外には原則として対応していない。

3.1.3 出力フォーマット

文字起こしされたファイルには，セグメント（行）の開始時間，終了時間，文字起こしテキストが含まれる。出力フォーマットは，以下のようなタブ区切りの TSV 形式か，カンマ区切りの CSV 形式から選択する。タイムスタンプが不要な場合は, テキストのみの TXT 形式での出力も可能である。Whisper.cpp は, VTT 形式，SRT 形式，JSON 形式にも対応しているが，現時点では本ツールでは対応していない。なお, 出力ファイルは, 原則として入力ファイルと同じフォ

ルダ内に生成される。

start	end	text
0	7600	“And so my fellow Americans ask not what your country can do for you,”
7600	10600	“ ask what you can do for your country.”

3.1.4 セグメント長

「セグメント長」は，1 セグメントに含まれる語数を意味する。設定は任意だが，指定しない場合は 1 セグメントが大幅に長くなる場合もあるため（図 1 は未指定），必要に応じて設定することが望ましい。特に，本ツールにおける検索文字列の計測は，1 セグメントに当該文字列が複数回生起している場合でも，1 度しかカウントされていない。セグメントを短くすることで，実態に近い値が得られることになる。

3.1.5 プロンプト

Whisper の基本的な文字起こしでは，“uh” や “um” などのフィラーや，“I have ... I have to” などの言い淀みは省略され，明らかな文法ミスもある程度修正されて文字起こしされることがある。本仕様は, 発話の内容把握には合理的だが, 言い淀みや言い誤りの調査には不都合である。こうした際，プロンプト機能が有効となる場合がある。

プロンプトに特定の語句を指定すると，それらの語句は原則として省略されずに文字起こしされる。プロンプト機能を使用した場合と，使用していない場合の結果を以下に示す。

［プロンプト機能を使用］	Um, they, they have to, they have to work after graduation.
［プロンプト機能を不使用］	They have to work after graduation.

（ICNALE Spoken: SM_JPN_PTJ1_024_A2_0）

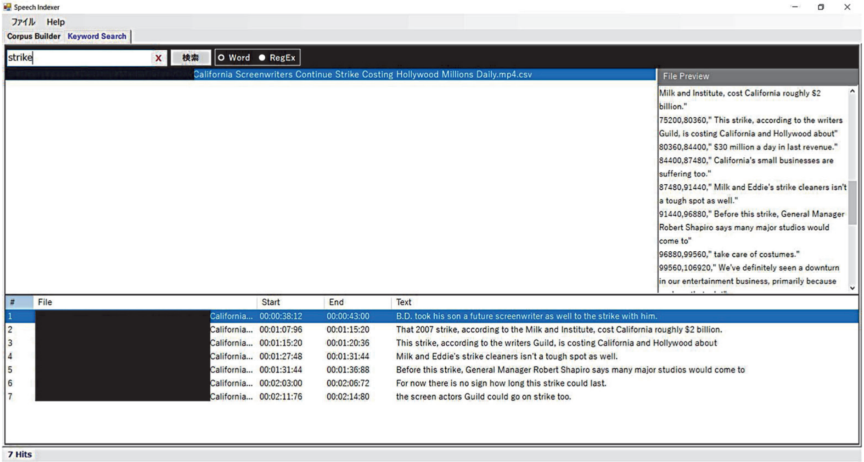
なお，デフォルトでは以下の語句をプロンプトとして設定しているが,「Prompt」フォルダ内にある “Filler_en.txt” を修正することで，指定する語句の設定が可能である。

[Umm, umm, Hmph, hmph, Um, um, Ah, ah, Uh, uh, Er, er, Oh, oh]

また，OpenAI のウェブサイトによると，プロンプト機能は人名や専門用語の認識精度の向上にも利用できるとされている。Speech Indexer ではその場合，プロンプト機能を有効にする際に，設定を「フィラー」ではなく「その他」を選択し，Others.txt ファイルに任意のプロンプトを記載する。ただし，4 節で示すように，プロンプト機能を有効にすると予期しない結果が得られることも多く，現時点ではやや実験的な機能と言える。

3.2 文字列検索

図 2 は，文字起こしされたファイル内の語句検索を行う画面である。通常のキーワード検索と同じ要領で，検索文字列をテキストボックスに入力して検索する。検索文字列を含むセグメントがあれば, 画面下の領域に「開始時間」「終了時間」「テキスト」が表示される。任意のセグメントをダブルクリックすることで，該当部分の音声または映像が別ウィンドウで再生される。



strike

○ Word ● RegEx

[redacted]California Screenwriters Continue Strike Costing Hollywood Millions Daily.mp4.csv

Milk and Institute, cost California roughly $2 billion."
75200,80360," This strike, according to the writers Guild, is costing California and Hollywood about"
80360,84400," $30 million a day in last revenue."
84400,87480," California's small businesses are suffering too."
87480,91440," Milk and Eddie's strike cleaners isn't a tough spot as well."
91440,96880," Before this strike, General Manager Robert Shapiro says many major studios would come to"
96880,99560," take care of costumes."
99560,106920," We've definitely seen a downturn in our entertainment business, primarily because

#	File	Start	End	Text
1	[redacted]California...	00:00:38:12	00:00:43:00	B.D. took his son a future screenwriter as well to the strike with him.
2	[redacted]California...	00:01:07:96	00:01:15:20	That 2007 strike, according to the Milk and Institute, cost California roughly $2 billion.
3	[redacted]California...	00:01:15:20	00:01:20:36	This strike, according to the writers Guild, is costing California and Hollywood about
4	[redacted]California...	00:01:27:48	00:01:31:44	Milk and Eddie's strike cleaners isn't a tough spot as well.
5	[redacted]California...	00:01:31:44	00:01:36:88	Before this strike, General Manager Robert Shapiro says many major studios would come to
6	[redacted]California...	00:02:03:00	00:02:06:72	For now there is no sign how long this strike could last.
7	[redacted]California...	00:02:11:76	00:02:14:80	the screen actors Guild could go on strike too.

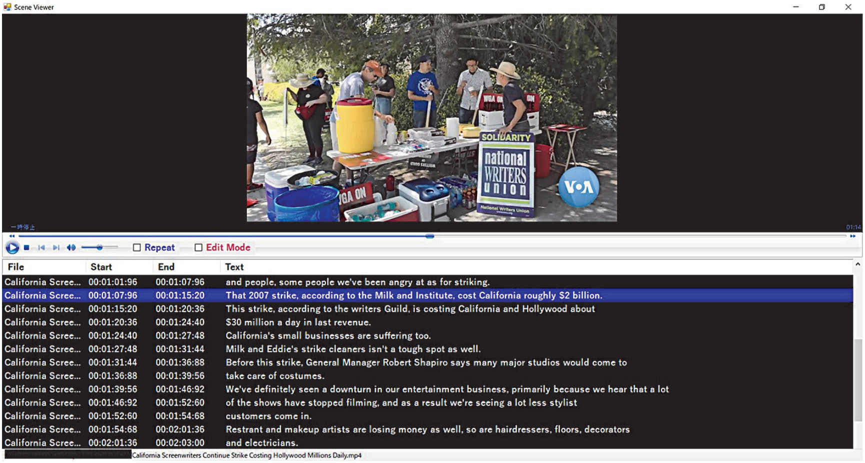
7 Hits

図 2. Speech Indexer の文字列検索画面

3.3 検索文字列該当箇所の再生

図 3 は，検索文字列がどのように発話されているかを確認する画面である。

画面上部に，音声ファイルの場合は波形が表示され，映像ファイルの場合は映像が音声と共に再生される。画面下部にはセグメント単位で文字起こしされたテキストが表示され，現在発話されているセグメントはハイライトされる。音声または映像の再生に合わせて，下部のテキスト表示画面も自動でスクロールする。特定のセグメントのリピート再生も可能である。



File	Start	End	Text
California Scree...	00:01:01:96	00:01:07:96	and people, some people we've been angry at as for striking.
California Scree...	00:01:07:96	00:01:15:20	That 2007 strike, according to the Milk and Institute, cost California roughly $2 billion.
California Scree...	00:01:15:20	00:01:20:36	This strike, according to the writers Guild, is costing California and Hollywood about
California Scree...	00:01:20:36	00:01:24:40	$30 million a day in last revenue.
California Scree...	00:01:24:40	00:01:27:48	California's small businesses are suffering too.
California Scree...	00:01:27:48	00:01:31:44	Milk and Eddie's strike cleaners isn't a tough spot as well.
California Scree...	00:01:31:44	00:01:36:88	Before this strike, General Manager Robert Shapiro says many major studios would come to
California Scree...	00:01:36:88	00:01:39:56	take care of costumes.
California Scree...	00:01:39:56	00:01:46:92	We've definitely seen a downturn in our entertainment business, primarily because we hear that a lot
California Scree...	00:01:46:92	00:01:52:60	of the shows have stopped filming, and as a result we're seeing a lot less stylist
California Scree...	00:01:52:60	00:01:54:68	customers come in.
California Scree...	00:01:54:68	00:02:01:36	Restrant and makeup artists are losing money as well, so are hairdressers, floors, decorators
California Scree...	00:02:01:36	00:02:03:00	and electricians.

[redacted] California Screenwriters Continue Strike Costing Hollywood Millions Daily.mp4

図 3. Speech Indexer の音声・映像確認画面

4. 学習者の英語音声認識精度

2.1 節で述べた通り，Whisper の英語の認識精度は約 95.5% である（Radford et al., 2022）。しかし，Radford らが評価に使用した「FLEURS Dataset」は基本的に英語母語話者の音声であり，英語学習者による英語の認識精度については十分に明らかにされていない。そこで本節では，ICNALE Spoken を使用し，日本人英語学習者の英語認識精度を調査した結果について報告する。

ICNALE（International Corpus Network of Asian Learners of English）は神戸大学の石川慎一郎氏が作成した英語学習者コーパスであり，日本をはじめ，アジア圏の国々における学習者の英文が収録されている。ICNALE Spoken はその中でも特に，学習者の発話を録音した音声ファイル，発話の様子を録画した映像ファイル，そしてそれらを手作業で文字起こししたテキストファイルを含む

音声・映像コーパスとなっている。

ICNLE Spoken には話者の CEFR レベルが付与されており，今回は A2，B1，B2 レベルから任意に 2 ファイルずつを選択し，調査に使用した。認識精度の評価については，音声認識評価において一般的に用いられる「単語誤り率：WER（Word Error Rate）」を利用した。Urban & Mehrotra（2023）によると，単語誤り率は以下のように計算できる。この時，I（Insertion）は「誤って追加された単語」，D（Deletion）は「検出されなかった単語」，S（Substitution）は「置き換えられた単語」，N は手作業で文字起こしされた正解単語数を意味する。

$$WER = \frac{I + D + S}{N} \times 100$$

I，D，S の具体例を図 4 に示す。上側の Human-labeled Transcript が手作業による文字起こし結果，下側の Speech Recognition Result が音声認識システムによって文字起こしされた結果である。

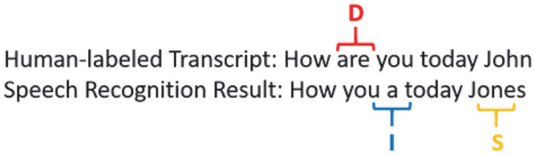

Human-labeled Transcript: How are you today John
Speech Recognition Result: How you a today Jones

図 4. 音声認識の誤り例（Urban & Mehrotra（2023）より）

表 2 は，使用したファイル名，話者の CEFR レベル，正解ファイルにおける単語数，及び WER を示したものである。WER は，プロンプト機能を使用した場合と，使用していない場合の両方を示している。モデルはいずれも LARGE を使用した。また，WER の計算には，「JiWER」と呼ばれる Python プ

表 2. 英語学習者による音声の認識結果

#	ファイル名	CEFRレベル	語数	WER （プロンプト無効）	WER （プロンプト有効）
1	SM_JPN_PTJ1_024_A2_0	A2	65	31.74%	80.95%
2	SM_JPN_PTJ1_050_A2_0	A2	37	123.53%	70.59%
3	SM_JPN_PTJ1_023_B1_1	B1	54	25.00%	28.84%
4	SM_JPN_PTJ1_046_B1_1	B1	49	40.00%	40.00%
5	SM_JPN_PTJ1_003_B2_0	B2	111	21.57%	18.62%
6	SM_JPN_PTJ1_005_B2_0	B2	75	5.48%	6.85%

ログラムを使用した。

一般に，WER は 10～20% 以下であることが望ましいとされている。B1 レベルと B2 レベルの認識精度については，「～_046_B1_1」がやや精度が低いものの，学習者の音声であることを考慮すると，概ね良好な結果であると思われる。特に「～_005_B2_0」において誤りとなった個所はカンマを含めた下線部のみであり，ほぼ正確に文字起こしされている（実際の正解データは“. Therefore, this experience”）。ただし，予想にやや反し，いずれもプロンプトの影響は特には見られなかった。

I agree with this opinion. There are three reasons. First, young people can learn something important about the relationship between elderly people and them. Second, young people can learn the importance of money through part-time job. Third, young people can use polite words, especially to elderly people. It is said that these days young people cannot know how to use polite words, therefore these experiences make them feel the importance of part-time job.

（ICNALE Spoken: SM_JPN_PTJ1_005_B2_0 ※プロンプト無効）

一方，A2 レベルでは WER が高いことに加え，プロンプトの有無による差も顕著となった。「～_024_A2_0」に関しては，プロンプトを有効にすることで，特定の音声パタンに過度に反応する挙動が見られた。具体的には，以下のように “um” が何度も繰り返されており，この音声ファイルの話者は確かにフィラーが散見されたが，ここまで極端なものではなかった。

Um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, um, Um, they, um, and, and, they, they want to, they need, they need to money.

（ICNALE Spoken: SM_JPN_PTJ1_024_A2_0 ※プロンプト有効）

また，「～_050_A2_0」については，話者のレベルよりも音声の録音環境の影響が大きかったと考えられる。当該ファイルは今回使用した音声の中で最も録音環境が悪く，メイン話者の発話が終わった後も背景で以下のような雑音が収録されてしまっていた。そのため，本来の発話の後に不要なテキストが追加されてしまっており，プロンプトの有無にかかわらず，WER が著しく高くなっ

ている。

Okay, this slide is your name, student number and part-time job number 1, E.T.J.1. Alright. Stop the button and then

（ICNALE Spoken: SM_JPN_PTJ1_050_A2_0 ※プロンプト無効）

以上，極めて小規模ではあるが，英語学習者の文字起こしにおける課題がいくつか浮き彫りとなった。一方，課題はあるものの，録音環境が整えられ，一定水準の流暢さで話す学習者であれば，決して低くない精度で文字起こしが行える可能性も示された。今後，英語母語話者の音声ではなく，ICNALE Spokenのような英語学習者の音声で訓練された音声認識システムが開発されれば，さらに精度の向上が見込まれる。より正確な音声認識が可能になれば，スピーキング学習や習熟度評価等への様々な利用も可能となり，今後のさらなる研究の発展が期待される。

5. まとめ

本稿では，音声認識システム「Whisper」・「Whisper.cpp」を利用した音声・映像コーパス構築ツール「Speech Indexer」について紹介した。Whisperは有用なプログラムであるが，プログラミングやコマンドライン操作に馴染みの薄い研究者にとっては使用するうえで少なからず技術的なハードルが存在すると考えられるため，本ツールの開発に至った。また，小規模ではあるが，英語学習者による英語の認識精度についても調査を行い，英語学習者による発話の文字起こしの可能性についても言及した。

音声認識システムは, 従来はコストの問題で困難となっていた音声･映像コーパスの開発を手軽に実践できる点において，コーパス言語学と極めて相性のいい技術であると考える。Speech Indexerが今後の音声・映像コーパス研究のさらなる進展に資することを願う。

注

Speech Indexerは筆者のウェブサイトにて公開中である。
https://www.setsunan.ac.jp/~corpus/SpeechIndexer.htm

引用文献

Hasebe, Y. (2015). Design and implementation of an online corpus of presentation transcripts of TED Talks. *Procedia: Social and Behavioral Sciences, 198*(24), 174–182.

Ishikawa, S. (2023). *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. https://arxiv.org/abs/2212.04356

Urban, E. & Mehrotra, N. (2023). Test accuracy of a custom speech model. *Microsoft Learn*. Retrieved December 11, 2023, from https://learn.microsoft.com/ja-jp/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio

参考ウェブサイト

FLEURS Dataset（https://paperswithcode.com/dataset/fleurs）（2023 年 8 月）
OpenAI『Introducing Whisper』（https://openai.com/research/whisper）（2023 年 8 月）
Whisper（https://github.com/openai/whisper）（2023 年 8 月）
Whisper.cpp（https://github.com/ggerganov/whisper.cpp）（2023 年 8 月）
JiWER（https://github.com/zszyellow/WER-in-python）（2023 年 11 月）

（後藤　一章　摂南大学　Email: goto@ilc.setsunan.ac.jp）

「BOOK REVIEWS」

Understanding the state-of-the-art of parallel corpus research in the world through *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* (Doval & Nieto, 2019)

Yasunori NISHINA

This book elucidates how parallel corpora can be compiled and utilized for contrastive studies, translation studies, and machine translation with the elucidation of the design and architecture of the parallel corpora and searching interface by the three parts of the book, including 16 chapters by the international researchers, preceded by the introduction by the two editors of the book.

The origin of parallel corpora dates back to 1971 regarding the Yugoslav Serbo-Croatian-English contrastive project by R. Filipovic. Since then, in the late 1980s, the Canadian Hansard Corpus of English and French Texts was compiled, followed by English-Norwegian Parallel Corpus (ENPC) and the English-Swedish Parallel Corpus (ESPC) in the 1990s. As Borin (2002: 1) has already stated in his book, "[i]n the last decade or so, parallel corpus linguistics has emerged as a distinct field of research within corpus linguistics, itself a fairly young discipline". Although the term 'parallel corpus linguistics' is not pervasive in the field, it has various possibilities to develop and expand the empirically-based language studies of two or more languages.

Within the 16 chapters as the main body of the contents in the book, the three chapters focus on the bilingual parallel corpora including an updated version of the ACTRES Parallel Corpus (P-ACTRES 2.0), Parallel Corpus of German and Spanish (PaGes), Corpus Linguístico da Universidade de Vigo (CLUVI), and Multidimensional Annotation of English-Spanish comparable and parallel texts for linguistic and computational applications (MULTINOT), while the four chapters deal with multilingual parallel corpora such as a part of the project Czech National Corpus (InterCorp), Valencian Corpus of Translated Literature (COVALT: the Corpus Valencià de Literatura Traduïda), the European Parliament Translation and Interpreting Corpus (EPTIC), the Parallel Electronic corpus of State Treaties (PEST), and a parallel/multilingual corpus consisting of literary texts translated from German to Basque (ALEUSKA). Most of the parallel corpora introduced in this book are text-based corpora, whilst two are multimodal (i.e. CLUVI and EPTIC). All parallel corpora consist of several text genres, text types and/or modes (e.g. written/spoken). At the same time, they are annotated at various levels.

The four chapters in the first part of the book are dedicated to how parallel and comparable corpora can usefully contribute to translation studies and contrastive linguistics, as presented in the title of the book, with the focus on some issues including the background/processing of parallel corpora and the recent developments of word-level alignment between two (or more) languages. L. Hareide's study adopted the applied use of two parallel corpora of the Norwegian Spanish Parallel Corpus (NSPC) and the first version of P-ACTRES as "comparable parallel corpora" to examine the gravitational pull hypothesis on the language pairs of Norwegian-Spanish and English-Spanish. The gravitational pull theory (Halverson, 2003, 2017) was proposed as a potential explanation for some general aspects of translated language. The underlying hypothesis is that highly salient linguistic items would be overrepresented in translational corpus data because they are more likely to be selected by translators. The NSPC and the English-Spanish P-ACTRES corpus are comparable to one another because they include similar-sized sub-corpora. With these corpora, the gravitational pull hypothesis was successfully tested, and the comparable parallel corpora method was proved useful in the Corpus-Based Translation Studies (CBTS). J. Marco's study, then, presented two case studies using the English-Catalan sub-corpus of COVALT in his chapter: the first study analyzes the translation of meal names with a parallel corpus as a main source of data, and the second analysis is on the construction *-ment* adverb + adjective as a supplementary source of data. R. Rabadán's study, on the other hand, provides her ideas about how parallel corpora can be useful and the details about the overview of resources, tools, applications, etc. This chapter firstly introduced the definition of parallel corpora, concepts of the usefulness/usability of parallel corpora, and several parallel corpora resources. Then, it reviewed the uses of parallel corpora and presented a needs analysis of parallel/multilingual corpora. Finally, it discussed whether to use the ready-made parallel corpora or build a brand-new parallel corpus, application for post-editing and assessment of translation, and useful strategies to start a new project using the parallel corpus. M. Volk's study also focuses on the annotation, alignment, and retrieval of the translation equivalents from parallel corpora, for example, giving some instances from the parallel corpus of English-German of film and TV subtitles.

The nine chapters of the book's second part present the ongoing parallel corpus project in European countries regarding the technical issues and aspects including corpus creation, annotation, and access. The current version of the InterCorp, a parallel corpus of Czech and 39 other languages compiled at Charles University in Prague, is elucidated by P. Čermák. Čermák mainly describes the size and structure of the corpus and the tools for the InterCorp, such as a corpus query tool Kontext, a text alignment editor tool InterText, and an automatic-built dictionaries of Czech-foreign languages Treq. The following chapter by I. Doval, S. Fernández Lanza, T. Jiménez, E. Liste Lamas and B. Lübke introduces the design and features of PaGes. This corpus considers the direct translation from one language to another, and does the translation direction to improve the accuracy of the contrastive analysis. The text processing, mark-up and metadata are detailed, followed by the alignment process. The functionalities of the searching tool interface for PaGes are also detailed, including the introduction of four search features, namely Fast search, User-friendly search, Multi-level search, and Display, followed by the server architecture and publishing data. In the chapter of A. Ferraresi and S. Bernardini's study, EPTIC compiled from EU parliament proceedings is detailed. The corpus includes 14 separate sub-corpora in which texts are aligned between texts and between text and video, enabling to display the actual delivery of the speech linked to the concordance lines. This is a multi-purpose corpus which currently includes three languages (English, French and Italian), from the two communication modes (spoken and written) and from the two translation modes (translated and interpreted).

X. G. Guinovart chapter then elucidates the description of the CLUVI corpus which is a sentence-level aligned parallel corpora compiled from the various genre texts of the nine languages such as Galician, Spanish, English, French, Portuguese, Catalan, Italian, Basque, and Latin. The twenty kinds of language combinations and domains are identified in the CLUVI. The corpus is human-annotated based on the TMX-based CLUVI Corpus XML specification and is available online. The VEIGA corpus is partly attached to multimedia data. The SensoGal corpus is also elucidated, which is an English-Galician parallel corpus semantically annotated using WordNet data. The following chapter by J. Lavid presents several issues that emerged from the annotation of the MULTINOT corpus, a parallel corpus of English and Spanish in both translation directions. In particular, most parts of this chapter detail the annotation of semantic, pragmatic, and discourse phenomena.

The chapter by M. Mikhailov, M. Santalahti and J. Souma describes PEST. PEST is a parallel corpus of treaties concluded between Russia-Finland, Finland-Sweden, and Sweden-Russia with a sub-corpus of international conventions in all the three languages as reference data. Among all, the structure of the sub-corpora of PEST is detailed, including the number of treaties over time, the topic of treaties, and so on. T. Molés-Cases and U. Oster's chapter introduces the architecture, compilation, and indexation of the part of COVALT corpus, a parallel corpus compiled from novels in English, French, and German as source languages and Spanish as a target language. Since the authors made it analyzable online with CQPweb, the sample query with CQPweb is also visually illustrated with some language syntax and CQP syntax options. H. Sanjurjo-González and M. Izquierdo describe P-ACTRES 2.0, an extended version of P-ACTRES 1.0 adding a new sub-corpus of original Spanish texts and their English translations. The workflow of building this corpus is well described such as compiling, formatting, aligning, tagging (with Treetagger), and indexing texts. For instance, formatting texts are separately explained such as cleaning texts, transforming them into XML format, validating XML files, and carrying out a sentence division. The usability of this corpus is also well presented through the instance of Spanish translational options of English construction *with* + NP + *-ing*. Finally, in Z. Sanz-Villar's chapter, the overview of the Basque corpora and the Aleuska corpus is described. The Aleuska corpus is a trilingual corpus of German literary texts with their translations of Basque and Spanish. The corpus design and compilation process using TAligner 3.0 are presented, followed by lemmatization and annotation at the POS level. Finally, the process of the extraction of Basque multi-word expressions (MWE) consisting of onomatopoeia and a verb is presented.

The three chapters in the third part of this book deal with the tools and applications of parallel corpora concisely from the three case studies. The study by P. Gamallo successfully presents that comparable corpora can be an alternative option to generate bilingual lexicons. The study introduces the two approaches of transitivity through intermediary dictionaries and refers to the similarity of bilingual cognates. It shows that accuracy is not much different when using a comparable corpus than when using a parallel corpus. The second chapter of the final part is the study by M. Garcia, M. García-Salido, and M. Alonso-Ramos about extracting bilingual collocation equivalents from parallel corpora. They utilize dependency parsing to extract the candidates of monolingual collocation and use the bilingual model of distributional semantics to identify the equivalents of the base and the collocate of the monolingual collocations. In the study procedure, the authors focus on lemmas instead of tokens regarding both the monolingual collocations and the distributional model. The cosine distance of the vectors of the two words is calculated to determine their semantic similarity. Using the parallel corpus of normalized and non-normalized French text messages, the last chapter by P. Goshal and X. Rao observes the efficacy of the two tools/approaches of *multivec* (multilingual word embeddings) and *moses* (character-based machine translation) for text normalization. The study concludes that *moses* and *multivec* differ in their preference for normalizing different categories. As a result, the authors suggest that the two approaches should be combined for a more robust precision and result in the normalization of shorthand forms in text messages.

As a whole, this book comprehensively introduces the development of parallel corpora and search tools in European countries at the present time, which makes it a valuable book for researchers, specialists, practitioners, and graduate students in the field. However, the volume of each part was noticeably different, and we would have liked to see improvements in this respect. In particular, since the overall focus was on the technical or technological aspects of parallel corpora, a few more specific case studies of CBTS and contrastive linguistics could have been introduced. For instance, a more concrete case study of computer lexicography would have conveyed more about the operational usefulness of parallel corpora and the potential for research development. To add, it would have been helpful to introduce parallel corpus studies of languages other than European languages, such as Asian languages.

Through this book, it can be also found that there are several limitations in the parallel corpus studies at the moment due to the limitation of bilingual/multilingual corpora available and the specificities of those corpora, as compared to the monolingual corpora. For instance, the size of EPTIC is still small, and only three languages are available out of 23. This situation is applied to other parallel corpora as well. The spoken parallel corpora are particularly scarce in the corpora of the certain pairs of languages (e.g. Japanese-English pair, see Nishina, 2023). The balance of translation direction is also important in parallel corpus studies. With some exceptions including MULTINOT, however, many parallel corpora are unidirectional.

In addition, the multimedia parallel corpora are scarce compared to the text-only corpora due to the scarcity of the translated multimedia materials. As X. G. Guinovart pointed out, its further development is strongly expected since the multimedia parallel corpora can be helpful for educational purposes, facilitating learners' autonomous language learning and providing meaningful information about ready-made subtitles to translators/practitioners.

Another issue in compiling parallel corpora is the copyright issue, whether most of the documents planned to be included in the corpora are publicly available. In the case of Japan, several Japanese-English parallel corpora are available, although a paid license agreement must be signed for the use of some of them (Nishina, 2023). For this reason, the development of parallel corpora incurs significant research costs.

Furthermore, the limited search tools and their functions available may be a problem. This is because the progress of tool development and its performance varies from language to language. There is a need for a unified platform on which researchers worldwide can share and collaborate to develop tools for bilingual/multilingual corpora. To add, as a parallel corpus linguist, I tackle the compilation of Japanese-English and English-Japanese parallel corpora and their searching interface, and have faced a similar situation as presented in this book; the sole researcher or a few members of a team feel that there is a limitation to developing and expanding the parallel corpus resources, interface, and researches. Co-operation with the international researchers/teams that have and provide the diverse expertise, skills, and resources is necessary for the near-future parallel corpora linguistics as easily expected. For instance, my team developed the lexical profiling system of the several Japanese-English parallel corpora available, named Parallel Link (https://www.parallellink.org/), and this tool enables any corpus users to extract the linguistic information they need quickly in terms of patterns, collocations, and instances with their translation equivalents. We hope to expand this tool into multilingual ones including, for example, Spanish. However, modern languages such as Spanish, French, and German are outside our field, and European researchers' assistance is required to succeed in this future project.

Finally, a general reference parallel corpus does not yet exist. Suppose there is a general reference parallel corpus. In that case, it can be utilized for the extraction of meaningful translation units, a compilation of bilingual dictionaries, empirical-based bilingual studies including translation studies/teachings, material developments, and language processing. As represented by British National Corpus (BNC) and Corpus of Contemporary American English (COCA), I sincerely hope that a general reference parallel corpus will be created in the near future.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 20K00692 and 23K00599. I would like to express my gratitude to the anonymous reviewer for his/her helpful comments and suggestions.

References

Borin, L. (2002). …and never the twain shall meet? In L. Borin (Ed.), *Parallel Corpora, Parallel Worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999* (pp. 1–43). Rodopi.

Halverson, S. (2003). The cognitive basis of translation universals. *Target, 15*(2), 197–241. https://doi.org/10.1075/target.15.2.02hal

Halverson, S. (2017). Gravitational pull in translation: Testing a revised model. In G. De Sutter, M. A. Lefer, and I. Delaere (Eds.), *Empirical Translation Studies: New Methods and Theoretical Traditions* (pp. 9–45). Mouton de Gruyter. https://doi.org/10.1515/9783110459586-002

Nishina, Y. (2023). *Aspects of Parallel Corpus Linguistics: From Monolingual Corpus Studies to Bilingual Corpus Studies (Society of Global Communication Studies of Kobe Gakuin University Research Series Vol.2)*. Kaitakusha.

（仁科　恭徳　神戸学院大学）

英語コーパス研究（第31号）
【2024年5月31日発行】

編集・発行　©2024 英語コーパス学会
〒464-8601 名古屋市千種区不老町
名古屋大学言語教育センター　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.31
2024

Research Articles

Software Introduction

Book Review

JAPAN ASSOCIATION FOR ENGLISH CORPUS STUDIES