Distant Co-occurrence Patterns of Connectives: a Corpus Study of Formulaicity in Japanese

Andrej BEKEŠ
University of Ljubljana, Slovenia
andrej.bekes@ff.uni-lj.si

Bor HODOŠČEK
Osaka University, Japan
hodoscek.bor.hmt@osaka-u.ac.jp

NISHINA Kikuko
Tokyo Institute of Technology, Japan
knishina@m06.itscom.net

ABEKAWA Takeshi
University of Tokyo, Japan
abekawa@p.u-tokyo.ac.jp

Abstract

Using corpus research methods, this study aims to establish whether there are two-item and, more generally, multi-item distant co-occurrence patterns of connectives in written Japanese, and further, to clarify the role these patterns play in discourse. The study is based on a hybrid corpus of written Japanese including Humanities and social science papers, Science and technology papers, and general written language data. The co-occurrence threshold was set at co-occurrence frequency > 10, PMI value > 2, and Dice coefficient > 0.01. The distribution of the observed co-occurring pairs differed according to the genre. Visualization of the connectivity potential of co-occurring pairs as directed graphs showed that these co-occurring pairs constitute longer co-occurrence chains which can be interpreted as ready-made co-occurrence patterns. Two-item and multi-item co-occurrence patterns are considered a type of Bourdieu's *habitus* and contribute to both discourse development and discourse prediction.

Keywords: connectives, distant co-occurrence, co-occurrence patterns, formulaic language, *habitus*, genre, directed graphs

Povzetek

Študija si z uporabo korpusnih metod prizadeva ugotoviti, ali obstajajo vzorci sistematičnega daljinskega sopojavljanja dveh veznikov oziroma, splošneje, več veznikov v pisni japonščini, ter nadalje pojasniti vlogo, ki jo ti vzorci izkazujejo v diskurzu. Študija temelji na hibridnem korpusu pisne japonščine, ki vključuje članke iz humanistike in družboslovja, znanosti in tehnologije ter splošnih pisnih besedil. Prag sopojavljanja dveh veznikov je bil nastavljen na frekvenco sopojavljanja > 10, vrednost PMI > 2 in koeficient Dice > 0,01. Porazdelitev opaženih sopojavljajočih se parov se je razlikovala glede na žanr. Vizualizacija povezljivosti sopojavljajočih se parov kot usmerjenih grafov je pokazala, da ti pari tvorijo daljše verige sopojavljajočih se veznikov, ki jih je mogoče razumeti kot že ustaljene vzorce sopojavljanja. Vzorce z dvema ali več sopojavljajočimi se vezniki lahko interpretiramo kot vrsto Bourdieujevega *habitusa*. Taki vzorci z vidika govorca prispevajo k razvoju diskurza, z vidika sogovorca pa k predvidevanju tega, kako se bo diskurz razvijal.

Ključne besede: vezniki, daljinsko sopojavljanje, vzorci sopojavljanja, formulaični jezik, habitus, žanr, usmerjeni grafi

Acta Linguistica Asiatica, 13(2), 2023.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.13.2.9-38

1 Introduction

1.1 Background

Ready-made patterns in discourse have been studied for a long time. The most typical of such patterns are syntax and collocations. Regarding the various structures in syntax, DeBeaugrande and Dressler (1981) point out that they act as an 'early warning system' for the listener/reader, facilitating processing.

Knowledge of formulaic language is also an important part of speakers' linguistic knowledge, and its study has a long tradition. Within this framework, more recently, Wray (2017) has focused on the systematic co-occurrence of various elements in linguistic data and discussed the important role this phenomenon has in the load reduction of language processing. There are also studies focusing on the Japanese language. One of these, Kaneyasu (2012), deals with systematically occurring morpheme sequences in Japanese conversation. Cognition-related findings from the study of formulaic expressions are important, but they are mostly limited to the patterns of occurrence of adjacent morphemes and their various functions. On the other hand, Ishiguro (2008, Chap. 10) discusses connectives from the perspective of 'strategic usage' (*senryakuteki shiyō*). These patterns can be observed at the discourse level and are based on systematically occurring chains of connectives.

1.2 Aims of the present study

This study is conceived as an exploratory study, focusing on the aforementioned 'strategic usage' patterns, and aims to investigate their reality and their relation to the role they play in discourse. Specifically, the aim is to investigate the systematic distant co-occurrence of connectives occurring at the beginning of sentences in a corpus of general and academic texts written in Japanese. For this purpose, the following research questions will be addressed.

RQ1: Is it possible to identify the most frequent and prominent patterns of distant co-occurrence of connectives in general and academic texts?

RQ2: If such identification is possible, is it then possible for multiple connectives to co-occur systematically?

RQ3: If systematic multiple co-occurrences are possible, what role do such co-occurrence patterns play in the actual discourse?

2 Previous research

There is a long tradition in Japan of studying the patterning of elements that are quite far apart syntactically. Minami's (1974) study of the hierarchical structure of Japanese clauses is a good example. Various original studies have also been conducted since then. Minami himself further statistically supported his earlier results in Minami (1993). Kudo (2000) corroborated the systematic nature of distant co-occurrence between sentence-initial adverbs and sentence-final modal expressions. This is an interesting result suggesting a kind of agreement phenomenon at the semantic level. This result by Kudo was further supported by Srdanović et al. (2009) in a large corpus of data.

Inspired by Noda (1995) and Kudo (2000), Bekeš (2008, Chap. 5; 2012) investigates the role of bracket structures formed by adverbs and co-occurring sentence-final modality expressions or adverbs and some *toritate* (focusing) particles and their role in discourse.

With the focus of the study shifting to connectives, the scope of analysis moves to the level of discourse. Within the framework of discourse research in Japan, there are many important findings. For example, Sakuma (2012, 2019) attempts to elucidate the role of connectives in the rhetorical structure of texts, based on studies such as Ichikawa (1978) and others, and on the establishment of detailed criteria for the identifying discourse units (i.e., written content paragraphs *bundan*, and spoken content paragraphs *wadan*).

There are also interesting studies on the systematic distant co-occurrence of connectives themselves. For example, Ishiguro (2008) points out the close relationship between connectives and sentence-final modality expressions (Ishiguro, 2008, Chap. 7). He also points to the possibility of systematic co-occurrence of multiple connectives and the existence of so-called ‘strategic usage’ in discourse development (Ishiguro, 2008, Chap. 10). Inspired by Ishiguro’s work, Wang Jinbo (2015a, 2015b) investigates the systematic co-occurrence of adversative (*gyakusetsu*) and additive (*junsetsu*) conjunctions in editorials and other genres, and classifies them according to their semantic properties. In order to elucidate the role they play in discourse, she further examines the correlation of such co-occurring pairs of connective expressions with their position in the discourse.

3 Methodology

3.1 Data

In this study the following data are used: the ‘Science and technology’ papers (hereafter shortened to ‘ST papers’), the ‘Humanities and social science’ papers (hereafter shortened to ‘HS papers’), and a partially modified BCCWJ*, representing the general use.[1] As connectives, listed in various dictionaries and other sources proved

[1] As for the science and technology papers, we independently collected data from the ‘Gengo shori gakkai’ (The Association for Natural Language Processing), ‘Doboku gakkai’ (Japan Society of Civil Engineers, ‘Nihon kagaku kai’ (The Chemical Society of Japan), ‘Nihon ikadaigaku kai’ (The Medical

to be insufficient data, the 523 connectives employed in Abekawa et al. (2020) were used in the analysis. Since their identification in sentences is very difficult in some cases, the analysis was limited to the most typical usage of connectives appearing at the beginning of sentences. The basic data of the relevant corpora used in this study are presented in Table 1.

Table 1: Basic data on corpora and connectives

Corpus	Total number of sentences (S)	No. of sentences that include one of the 523 connectives	Percentage of the total (S)
BCCWJ*	3,204,314	581,461	18.1%
HS papers	447,645	130,601	29.2%
ST papers	1,182,181	169,872	14.45%

In addition to these corpora, a small corpus consisting of 300 Asahi Shimbun editorials and opinion articles was used for validation.

3.2 Method of analysis

The extraction of co-occurrence patterns relied on two measures of association, i.e., the PMI[2] (pointwise mutual information) and Dice coefficient[3], both of which are used

Association of Nippon Medical School), 'Kankyō shigen kōgaku kai' (The Resources Processing Society of Japan) and 'Denki gakkai' (The Institute of Electrical Engineers of Japan). In the case of the journal of The Association for Natural Language Processing, data include papers and proceedings of annual conferences, while in the case of other societies' journals, papers alone were collected. The total number of papers and proceedings included in this corpus is 4,865.

Humanities and social sciences papers. From J-STAGE, a general academic e-journal site, we independently collected up to 20 papers per each relevant academic society from the academic journals they publish by specifying the search field as Humanities and Social Sciences. The total number of papers collected in this way is 1,508.

BCCWJ*. Based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a corpus built by the National Institute for Japanese Language and Linguistics (NINJAL) in order to provide a comprehensive picture of the written modern Japanese language use, and divided into several media types/genres. BCCWJ* is a sub-corpus of BCCWJ. In order to limit the data only to those written prose texts that are considered to have been thoroughly proofread, material taken from 'Yahoo! Chiebukuro', 'Yahoo! blog', Diet Minutes, and, for genre reasons, poetry, were excluded. See also Abekawa et. al. (2020).

[2] A statistical measure of association that compares the probability of two events occurring together to the probability of them occurring independently. For two outcomes of random variables x and y, PMI is defined as PMI(x,y)=$\log_2$P(x,y)/P(x)P(y), where P(x,y) is the probability of x and y occurring together and P(x), P(y) are the probabilities of x and y occurring independently. The higher the value, the stronger the degree of association between x and y.

[3] Another statistical measure of association, the Dice coefficient is defined as Dice(x,y) = 2f(xy)/(f(x) + f(y), where f(x) and f(y) are the frequencies of words x and y, and f(xy) is the frequency of words x

to indicate the degree of association of co-occurring items used in traditional corpus studies (see Kolesnikova, 2016, Petrovic et al., 2006). The semantic relations underlying distant co-occurrence at the discourse level differ from those found in collocation studies which focus on local relations within a sentence. Therefore, at this stage of the study, it was empirically decided to use both measures together. Co-occurrence extraction was restricted to pairs of connected expressions that co-occurred within a range of one to four sentences apart in context. No attempt was made to extract more than two multiple connectives (n-grams), as the larger the number, the less reliable the results, due to the increased distance between individual items. Instead, typical co-occurrence pairs of the extracted connectives were considered as arcs in a directed graph, concatenated into longer co-occurrence chains, and further validated on the basis of actual data.

4 Analysis

4.1 Co-occurrence and genre of connectives

The first conjunctive expression appearing in co-occurrence is labeled as X and the second as Y. Their co-occurrence frequency is denoted as f(XY). Following Sakuma (2012, 2019), discourse units (content-based paragraphs) are referred to as *dan*. For easier recognition of co-occurrences, *dan* is assumed to consist of one or more sentences (S). A content-based paragraph realized by two co-occurring connectives can be denoted as follows.

(1) ［dan_0＝S_0］ -X- ［dan_1＝S_1］ -Y- ［dan_2＝S_2］

In other words, two co-occurring connectives represent a relationship between three *dan* content paragraphs. The relationship may be parallel or hierarchical. Following the custom in corpus analysis, to eliminate rare and thus considered atypical cases of co-occurrence, only cases with frequency f(XY) > 10 are included in the analysis. On the other hand, in order to widen the range of potential co-occurrence candidates, the threshold value of PMI is set to PMI > 2, which is lower than the usual threshold value of PMI > 4, customarily used in corpus analysis (cf. Petrovic, 2006). And finally, an additional condition, i.e., Dice coefficient > 0.01, is added to the co-occurrence recognition criteria, to compensate for the tendency of PMI to be higher for low-frequency co-occurrences. The co-occurrences meeting these criteria are tabulated in Table 2 below:

and *y* occurring together. The higher the value, the stronger the degree of association between *x* and *y*.

Table 2: Aggregation of co-occurrences in the three corpora

Corpus	Total number of sentences (S)	No. of sentences that include one of the 523 connectives	Number of co-occurrences of X, Y: (K as % of S)	No. of X,Y such that f(XY)>10, PMI>2, DICE>0,01 (C as % of K)
BCCWJ*	3,204,314	581,461	2708 (0.47%)	87 (3.21%)
HS papers	447,645	130,601	857 (0.66%)	202 (23.57%)
ST papers	1,182,181	169,872	925 (0.54%)	181 (19.57%)

It is clear from the table that the percentage of connectives that co-occur with other connectives within a certain range (four sentences) is low, around 0,5%, regardless of the corpus. This means that in the corpora studied, the proportion of connectives that explicitly indicate the relationship between two *dan* content paragraphs is low. On the other hand, the proportion of co-occurrences C that exceed the PMI value and Dice coefficient thresholds among all co-occurrences K in their respective corpora is significantly lower in BCCWJ* at 3.21%, compared to 23.57% in 'Humanities and social sciences papers' and 19.57% in 'Science and technology papers'. This means that in both academic corpora, the relationship between the three *dan* content paragraphs shown in (1) is significantly more likely to be systematically made explicit by connectives than in BCCWJ*.

4.2 Top 20 co-occurrence examples in PPM and PMI

To further clarify the systematic co-occurrence of connectives by genre, let us compare the top 20 co-occurrences in PPM (occurrences per million cases) and PMI respectively.

Table 3: Top 20 co-occurrences based on PPM: BCCWJ* and academic papers data

BCCWJ*					HS papers					ST papers				
Connective X	Connective Y	Freq.	PPM	PMI	Connective X	Connective Y	Freq.	PPM	PMI	Connective X	Connective Y	Freq.	PPM	PMI
mata	*sarani*	1536	381.4	4.1	*mata*	*sarani*	393	877.9	5.2	*shikashi*	*sokode*	843	713.1	5.9
shikashi	*mata*	896	222.5	**1.2**	*shikashi*	*mata*	380	848.9	3.8	*shikashi*	*mata*	584	494.0	3.7
mata	*mata*	825	204.9	**1.4**	*tatoeba*	*mata*	308	688.0	5.1	*mata*	*sarani*	523	442.4	4.7
mata	*shikashi*	816	202.6	**1.0**	*mata*	*shikashi*	298	665.7	3.4	*mata*	*mata*	484	409.4	2.2
shikashi	*soshite*	747	185.5	**1.3**	*shikashi*	*sokode*	229	511.6	5.2	*tatoeba*	*mata*	440	372.2	3.6
tatoeba	*mata*	693	172.1	2.9	*tashikani*	*shikashi*	204	455.7	7.2	*mata*	*sokode*	438	370.5	3.7
shikashi	*sokode*	689	171.1	2.7	*shikashi*	*tsumari*	194	433.4	4.5	*mata*	*tatoeba*	429	362.9	3.5
mata	*nao*	627	155.7	3.2	*shikashi*	*tatoeba*	193	431.1	4.5	*mazu*	*tsugini*	413	349.4	7.1
soshite	*shikashi*	624	155.0	**1.1**	*mata*	*soshite*	183	408.8	3.6	*mata*	*shikashi*	371	313.8	3.1
tashikani	*shikashi*	550	136.6	4.0	*mata*	*konoyōni*	178	397.6	4.6	*mata*	*nao*	346	292.7	3.9
mochiron	*shikashi*	542	134.6	3.0	*mata*	*mata*	177	395.4	2.5	*kokode*	*mata*	336	284.2	3.6
sarani	*mata*	492	122.2	2.4	*shikashi*	*soshite*	168	375.3	3.6	*tadashi*	*mata*	322	272.4	4.1
shikashi	*tatoeba*	477	118.4	2.1	*soshite*	*shikashi*	156	348.5	3.5	*shikashi*	*tatoeba*	314	265.6	4.3
nao	*mata*	471	117.0	2.8	*soshite*	*mata*	153	341.8	3.3	*nao*	*mata*	313	264.8	3.8
shikashi	*tsumari*	440	109.3	2.1	*shikashi*	*sonotame*	149	332.9	5.0	*mata*	*shitagatte*	266	225.0	3.9
mata	*konoyōna*	434	107.8	3.0	*tsumari*	*mata*	147	328.4	4.0	*sokode*	*mata*	254	214.9	2.9
mata	*ippō*	409	101.6	2.9	*mata*	*shitagatte*	143	319.4	4.5	*shikashi*	*sonotame*	248	209.8	5.5
soshite	*soshite*	408	101.3	**1.1**	*mata*	*tsumari*	139	310.5	3.9	*tatoeba*	*shikashi*	238	201.3	3.9
mata	*konoyōni*	390	96.8	3.2	*mata*	*tatoeba*	130	290.4	3.8	*mata*	*ippō*	215	181.9	3.5
tatoeba	*shikashi*	384	95.4	**1.8**	*mata*	*nao*	129	288.2	4.3	*ippō*	*mata*	212	179.3	3.5

It is immediately apparent from Table 3 that a high or low co-occurrence frequency expressed as PPM does not necessarily correlate with a high or low PMI value, with the *mata → shikashi* combination in BCCWJ* being a striking example of this. In Table 3, PMI values below the threshold value used here, i.e., < 2, observed in BCCWJ*, are highlighted with boldface. In both academic corpora, however, all PMI values are above the threshold, and the correlation between PPM and PMI values appears to be more consistent.

On the other hand, the PMI of co-occurrence of *shikashi* (however) → *sokode* (therefore) is consistently high, regardless of genre, with a PMI value of 2.7 for BCCWJ*, 5.2 for Humanities and social sciences papers and 5.9 for Science and technology papers. The PMI values of this co-occurrence are significantly higher in the academic corpus than in the BCCWJ*. This can be attributed to the fact that the range of observed combinations of connectives is more limited and more formulaic in the academic corpora than in the BCCWJ*. This is also clearly visible in the directed graphs visualization discussed in section 4.3. In Table 3, all co-occurrences in which the PMI value is above the threshold, but co-occurs with a high PPM, regardless of genre, could intuitively be considered as established co-occurrence patterns.

Finally, let us have a look at the use of *shikashi* (however) → *sokode* (therefore), one of the cases with the highest frequency expressed as PPM, using an editorial as an example.

(2) 日本郵政をめぐって、民主党政権は株式売却の凍結法を成立させる一方、郵政改革を抜本的に見直す法案を昨年の通常国会に出した。その成立をみて資産売却を解禁する段取りを描く。しかし、見直し法案には、小泉政権下で民営化を断行した自民党など野党の反対が根強い。そこで、復興のための増税圧縮に絡めて成立を急ごうという思惑が垣間見える。
（朝日新聞 2011 年 09 月 17 日、朝刊）

Nippon'yūsei o megutte, Minshutō seiken wa kabushiki baikyaku no tōketsu-hō o seiritsu sa seru ippō, yūsei kaikaku o bappon-teki ni minaosu hōan o sakunen no tsūjō kokkai ni dashita. Sono seiritsu o mite shisan baikyaku o kaikin suru dandori o kaku. Shikashi, minaoshi hōan ni wa, Koizumi seiken-ka de min'ei-ka o dankō shita Jimintō nado yatō no hantai ga nedzuyoi. Sokode, fukkō no tame no zōzei asshuku ni karamete seiritsu o isogou toiu omowaku ga kaimamieru.
(Asahi shinbun, 2011 nen 09 tsuki 17 nichi, chōkan)

With regard to Japan Post, the Democratic Party of Japan (DPJ) government passed a law freezing the sale of shares, while a bill to fundamentally review postal reform was submitted to the ordinary session of the Diet last year. The government will draw up arrangements to lift the ban on asset sales when the bill is passed. However (*shikashi)*, the revised bill faces strong opposition from opposition parties, including the Liberal Democratic Party (LDP), which had decisively privatized the postal service under the Koizumi administration. Therefore (*sokode*), there are glimpses of a desire to hasten its passage by tying it to the compression of tax hikes for reconstruction.

(Asahi Shimbun 17 Sep 2011, morning edition, editorial)

In example (2), the PPM of the co-occurring pair *shikashi* (however) → *sokode* (therefore) in different corpora is as follows: BCCWJ* 171.1; HS 511.6; ST 713.1. PPM is significantly higher in academic data (about four times higher than BCCWJ* in 'Science and Technology papers' and about three times higher in 'Humanities and social sciences papers'). Example (2) is an example of a development in which, in a given situation, an adversative conjunctive expression such as *shikashi* (however) introduces an inconvenient situation and *sokode* (therefore) introduces a response to the situation (here, it is, 'hastening of the passage of bill'. This co-occurrence pattern is discussed in detail in Wang (2015a,b).

Next, let us look at the top 20 co-occurrences with the highest PMI values.

Table 4: Top 20 co-occurrences based on PMI values: BCCWJ* and academic papers data

BCCWJ*					HS papers					ST papers				
Connective X	Connective Y	Freq.	PPM	PMI	Connective X	Connective Y	Freq.	PPM	PMI	Connective X	Connective Y	Freq.	PPM	PMI
dainiwa	*daisanwa*	39	**9.7**	12.5	*hitotsumewa*	*futatsumewa*	11	24.6	14.4	*hitotsumewa*	*futatsumewa*	13	11.0	14.9
daisan'ni	*daiyon'ni*	30	**7.4**	12.3	*zenshawa*	*kōshawa*	18	40.2	13.1	*daiichini*	*dainini*	16	13.5	14.0
daiichiwa	*dainiwa*	39	**9.7**	12.2	*daisan'ni*	*daiyon'ni*	16	35.7	12.3	*zenshawa*	*kōshawa*	30	25.4	13.2
zenshawa	*kōshawa*	44	10.9	12.1	*hitotsuwa*	*mōhitotsuwa*	25	55.8	12.0	*hitotsuwa*	*mōhitotsuwa*	34	28.8	13.1
dainini	*daisan'ni*	81	20.1	11.3	*dainini*	*daisan'ni*	55	122.9	11.1	*hitotsuwa*	*mōhitotsu*	14	11.8	11.1
daiichini	*dainini*	63	15.6	10.8	*daiichini*	*dainini*	86	192.1	11.0	*tsugini*	*saigoni*	126	106.6	7.1
hitotsuwa	*mōhitotsuwa*	123	30.5	10.8	*hitotsuwa*	*mōhitotsu*	15	33.5	10.4	*mazu*	*tsugini*	413	349.4	7.1
hitotsuwa	*futatsumewa*	21	**5.2**	9.7	*saigoni*	*daiichini*	13	29.0	7.8	*sonotaa*	*sonotaa*	14	11.8	6.8
uchi	*uchi*	21	**5.2**	9.4	*mazu*	*tsuide*	11	24.6	7.6	*dewa*	*dewa*	38	32.1	6.5
hitotsuwa	*mōhitotsu*	67	16.6	8.1	*tsugini*	*saigoni*	41	91.6	7.5	*shikashinagara*	*sokode*	134	113.3	6.0
hitotsuwa	*zenshawa*	12	**3.0**	7.6	*mazu*	*tsugini*	102	227.9	7.5	*sokode*	*gutaitekiniwa*	133	112.5	6.0
ika	*ika*	376	93.4	7.2	*tashikani*	*shikashi*	204	455.7	7.2	*kekka*	*kekka*	27	22.8	5.9
ika	*konobaainioite*	55	13.7	7.1	*ikadewa*	*mazu*	20	44.7	7.1	*shikashi*	*sokode*	843	713.1	5.9
konobaainioite	*daiichi*	33	**8.2**	6.6	*ikadewa*	*tsugini*	14	31.3	7.1	*mazu*	*soshite*	116	98.1	5.8
daiichi	*konobaainioite*	31	**7.7**	6.5	*soredewa*	*mazu*	19	42.4	7.0	*kokodewa*	*mazu*	48	40.6	5.7
aruiwa	*izureniseyo*	20	**5.0**	6.3	*mazu*	*tsudzuite*	12	26.8	7.0	*koremade*	*shikashi*	61	51.6	5.5
daiichi	*dainini*	27	**6.7**	6.3	*daga*	*toiunomo*	18	40.2	6.7	*soshite*	*saigoni*	29	24.5	5.5
daiichi	*daiichi*	153	38.0	6.2	*mochiron*	*daga*	24	53.6	6.6	*shikashi*	*sonotame*	248	209.8	5.5
nande[a]	*nan**de***[a]	17	**4.2**	6.1	*mochiron*	*shikashinagara*	18	40.2	6.6	*shikashinagara*	*sonotame*	34	28.8	5.4
de	*de*	29	**7.2**	6.0	*mazu*	*sonōede*	12	26.8	6.5	*shikashinagara*	*tokuni*	20	16.9	5.4

[a] *nande* is contraction of *nanode*

At high PMI values, once the first connective X is selected, the subsequent connective Y is predictable to a significant degree. In Table 4, the co-occurrence frequencies (number of co-occurrences) in the academic paper data are all greater than 10. However, in BCCWJ*, which has the largest amount of data, 12 out of 20 co-occurrences have PPM values below 10, while their PMI values are high, some even very high (> 10). The co-occurrence with the lowest PPM value but still considerably high PMI (> 9) is *hitotsuwa* (one) → *zenshawa* (the previous one) and the second lowest is *aruiwa* (or) → *izureniseyo* (in any case).

In all three corpora, some co-occurrences with relatively low frequencies but PPM values of 10 or more are also considered highly predictable because of their rather high PMI values: in BCCWJ* there are 8 such occurrences out of the top 20 co-occurrences, in Humanities and social sciences papers 15 out of the top 20 in and in Science and technology papers, 9 out of the top 20. Most of these high PMI co-occurrences seem to correspond to examples classified as 'enumerations (*seiri-rekkyo*)' in Ishiguro (2008), such as *hitotsuwa* (one) → *mōhitotsuwa* (the other) above. Although only used in a limited number of contexts, these are examples with a very high degree of formulaicity.

Let us now have a look at an example of *hitotsuwa* (one) → *mōhitotsuwa* (the other) from an opinion article.

(3) 環境省が原発規制の元締となることには、**二つの意味**がある。一つは、国策としての原発推進の終わりだ。＜……＞
もう一つは、原発を支えてきた安全神話の終わりだ。
（朝日新聞2011年08月16日、朝刊、オピニオン）

Kankyōshō gap genpatsu kisei no motojime to naru koto ni wa, ***futatsu no imi*** *ga aru. Hitotsu wa, kokusaku to shite no genpatsu suishin no owarida. <……>*
Mōhitotsu wa, genpatsu o sasaete kita anzen shinwa no owarida.
(Asahi shinbun 2011 nen 08 tsuki 16 nichi, chōkan, Opinion)

The Ministry of the Environment becoming the prime regulator of nuclear power has **two implications**. One (*hitotsu*) is the end of the promotion of nuclear power as a national policy. <A detailed description spanning over six sentences, follows.> The other (*mōhitotsu*) is the end of the safety myth that has underpinned nuclear power.

(Asahi Shimbun, 16 Aug 2011, morning edition, Opinion).

In (3), the co-occurring pair *hitotsuwa* (one) → *mōhitotsuwa* (the other) has the PMI value as follows: BCCWJ* 10.4; HS 10.4; ST 11.1. This is a clear and relatively common example of 'enumeration'. The pair of connectives is introduced by the cataphoric reference of '**two implications**', and the relationship between the first relatively long *dan* content paragraph and the subsequent *dan* content paragraph is also clearly indicated by both connectives. In this case, the high PMI value also implies a developed formulaicity of the co-occurring pair.

Three other interesting co-occurrence patterns are found in the Humanities and social sciences data. These are *tashikani* (certainly) → *shikashi* (however), *mochiron* (of course) → *daga* (but), and *mochiron* (of course) → *shikashinagara* (however). All three have fairly similar meanings, and the sequence is signaling a rhetoric pattern, i.e., 'acceptance of (collocutor's) proposition, followed by an alternative proposal'. This pattern appears to be used as a strategy to express cautious disagreement or for introducing additional alternatives. On the other hand, such co-occurrence examples do not seem to be used very often in the Science and technology papers data, but in Humanities and social sciences data, the PMI values and PPM of these cases are twice as high as in BCCWJ*. For example, in co-occurrence pair *mochiron* (of course) → *daga* (but) the PMI values are as follows: BCCWJ* 2.64; HS 6.6; ST N/A. Let us have a look at an example, again from an editorial.

(4) 素人である検察審査会の審査員や裁判員に正しい判断ができるのか、という声はくすぶる。
もちろん、絶対に間違えないとは言わない。だが国民の能力をうんぬんする以前に、専門家の手で正しい証拠が隠されたり、不当な誘導がされたりすることが、誤った結論をもたらす。…
（朝日新聞 2011 年 12 月 18 日、朝刊、社説）

Shirōto dearu kens Atsu shinsa-kai no shinsa-in ya saiban-in ni tadashī handan ga dekiru no ka, to iu koe wa kusuburu.
Mochiron, zettai ni machigaenai to wa iwanai. Daga kokumin no nōryoku o un'nun suru izen ni, senmonka no te de tadashī shōko ga kakusa re tari, futōna yūdō ga sa re tari suru koto ga, ayamatta ketsuron o motarasu.…
(Asahi shinbun 2011 nen 12 tsuki 18 nichi, chōkan, shasetsu)

There are persistent voices asking whether lay prosecution jurors and lay jurors are capable of making the right decision.
Of course (*mochiron*), this is not to say that they will never make a mistake. But (*daga*) before we can say anything about the competence of the public, the fact that the right evidence is hidden or improperly guided by experts leads to erroneous conclusions. ...

(Asahi Shimbun, 18 Dec 2011, morning edition, editorial)

In (4), first the doubt about the ability of lay jurors is presented. Of course (*mochiron*) then introduces the acceptance of the possibility that they could make mistakes. The argument is then countered in the next sentence, introduced by but (*daga*), which emphasizes the fact that it is actually the specialists who in many cases mishandle the evidence, leading to erroneous conclusions.

On the other hand, there are co-occurrence pairs in the top 20 PMI values that are only found in the Science and technology papers. These are the most frequently co-occurring pairs among the top 20 PMI co-occurrences, i.e., *shikashi* (however) → *sokode* (therefore), and its also frequent alternative *shikashinagara* (however) → *sokode* (therefore). As we have already seen, the first of the pairs, *shikashi* (however) → *sokode* (therefore) is relatively frequent and found in all corpora in the top 20 PPM co-occurrences.

In addition, there are two other frequent co-occurrence pairs including *shikashi* (however) and *shikashinagara* (however) as the first member, i.e., *shikashi* (however) → *sono tame* (therefore) and *shikashinagara* (however) → *sono tame* (therefore). Both pairs introduce 'another aspect of a given situation followed by its consequences'.

It is not only between the BCCWJ* data and the academic corpora that significant differences in the distribution of the high-frequency co-occurrence pairs can be observed. Interestingly, such differences are also found between the Humanities and social sciences data and Science and technology data. This suggests that differences arise not only between the genres such as general and academic use but also between different academic disciplines.

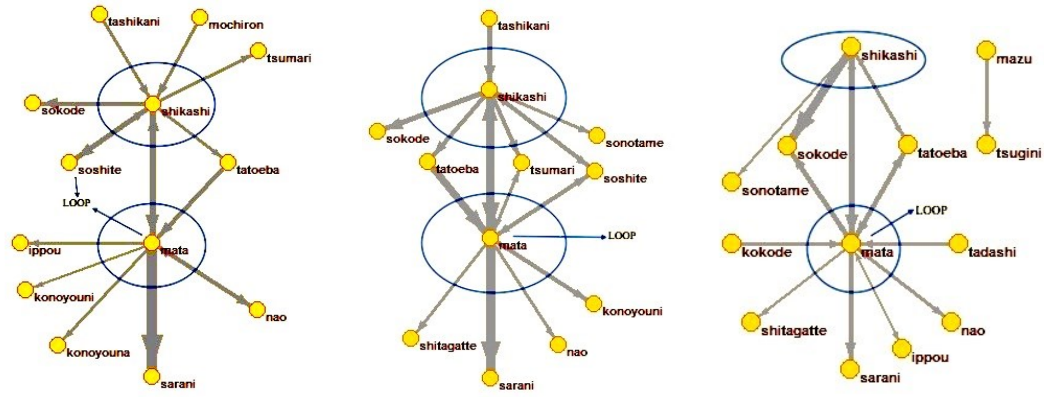
On the basis of Tables 3 and 4, we can conclude that when the co-occurring pair of connectives meets the threshold values for frequency, PMI, and additionally, the Dice coefficient, those pairs displaying either high frequency and medium PMI values, or relatively low frequency but high PMI values, can be considered intuitively as having developed into formulaic pairs. This finding allows RQ1 to be answered in the affirmative.

4.3 Identifying longer co-occurrence chains by directed graphs

As indicated in (1), the role of connectives is to explicitly indicate semantic relations between *dan* content paragraphs in a sequence of *dan* content paragraphs. Thus, behind the sequence of connectives, there is in fact a sequence of *dan* content paragraphs in the larger discourse unit that contains them. As the discourse unfolds in time (or space in the case of a written text), the pairs of connectives can be thought of as directed graphs. The connectives X and Y are nodes in the graph and 'X→Y' is the direction from X to Y. As one connective may co-occur with a number of other connectives in a context, co-occurring pairs can be linked into even larger chains. By identifying those chains of connectives that occur systematically in the context, i.e., Ishiguro's 'strategic usage' patterns, conjunctive relations behind them can be explored.

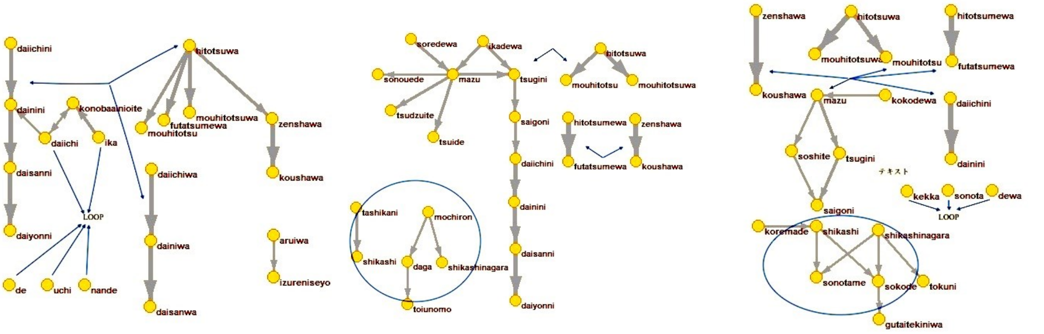
In the rest of this section, we discuss the possibility of identifying potential chains of connectives by representing the identified co-occurrence pairs as directed graphs. For this purpose, we use 'Pajek', a graph exploration software (see de Nooy et al., 2005).

Based on the co-occurrence data in Tables 3 and 4, visualization of potential, longer co-occurrence chains containing multiple co-occurring pairs is shown in Figure 1a and Figure 1b. Depending on one's point of view, this visualization can be interpreted as the potential knowledge of the use of connectives in a given community of language users (including a community of experts), i.e. it represents an aspect of what de Saussure (1916/1966) calls *langue*, or what Bourdieu (1991, 1994) calls *habitus.*



PPM BCCWJ* PPM Hum. and soc. science PPM Science and technology

Figure 1a: Top 20 PPMs - directed graph with concatenated co-occurrence sequences



PMI BCCWJ* PMI Hum. and soc. science PMI Science and technology

Figure 1b: Top 20 PMI-directed graph with concatenated co-occurrence sequences

Figure 1a and Figure 1b visualize all the top 20 PPM and PMI co-occurrence pairs by integrating them into directed graphs. As the graphs in Figure 1a and Figure 1b reflect co-occurrence frequencies (PPM) or PMI values based on corpus data, they do not show actual linkage relations in specific contexts, but only potential co-occurrence patterns of language use within the general user community and within specific scientific communities.

In the graphs, the thickness of the arcs connecting the nodes is proportional to the frequency (PPM) of co-occurrence or to the PMI value. In Figure 1a, based on the high frequency (PPM) co-occurrences shown in Table 3, in all three genres, the bidirectionally connected *shikashi* (however) and *mata* (also) (in Figure 1a marked by circles) form a central pair of nodes with which various relations of connectives can be formed. Not only in terms of their position but also in terms of the frequency of co-occurrence, the connectives can be linked into longer chains around *shikashi* (however) and *mata* (also). However, the details vary from genre to genre.

On the other hand, for co-occurrences with high PMI values (Figure 1b), two types of co-occurrence sequences emerge. One type is potential sequences (marked with arrows) that clearly represent Ishiguro’s multiple ‘enumerating’ strategy patterns mentioned above. The thickness of the arcs reflects the predictability of co-occurrence, thicker arcs representing higher PMI values and thus higher predictability of co-occurrence. The other type of patterns that emerge clearly in both academic corpora are the patterns (marked by circles) that appear to be used to develop the argumentation, although the PMI values involved in them are somewhat lower. In Humanities and social sciences data, this is the pattern *mochiron* (of course) → *daga* (but), discussed in section 4.2, and in Science and technology data, the pattern of *shikashi* (however) → *sokode* (therefore), also mentioned in 4.2.

In several places in Figure 1a and Figure 1b, there is a node called ‘LOOP’. This node represents cases where, according to the specification of the ‘Pajek’ software used, one conjunctive expression co-occurs repeatedly with itself, such as *de* (and) → *de* (and), etc. LOOP is not combined with other connectives with a very high PPM (frequency) or high PMI value (predictability), except for *mata* (also).

In this section, we have looked at the co-occurrence of the most frequently occurring connectives and potential argumentation patterns based on them. On the basis of the above findings, the answer to RQ2 can be regarded as affirmative.

These potential patterns, which, with the help of visualization, suggest strategies for the development of longer discourse segments, can be said to be the realization of the linking potential possessed by the two-item co-occurrence examples presented in Tables 3 and 4.

Let us now look at the potential chains containing two or more connectives that we obtained from Figure 1a and Figure 1b, i.e., chains based on the high co-occurring frequency (PPM) and high PMI values. In order to clarify the relationship between the potential chains including two or connectives and the *dan* content paragraphs, having the chain schema in (1) as a starting point, a more general form of the chain of connectives is shown in (5).

(5) $\boxed{dan_0}\text{-}X_1\text{-}\boxed{dan_1}\text{-}X_2\text{-}\boxed{dan_2}\ldots\boxed{dan_{n-1}}\text{-}X_n\text{-}\boxed{dan_n}$

In general, *dan* content paragraphs are not necessarily sentences but can consist of a group of sentences. The semantic relationship between dan_i and dan_{i+1} generally need not be explicitly indicated with the connective X_i. The unexpressed X_i is denoted as Ø for convenience, as a ‘fill-in’. In fact, examples of conjunctive relations between sentences that are not explicitly signaled (i.e., are realized as ‘Ø’), account for about two-thirds or more of all co-occurrences identified in all three corpora examined in this study.

As it is very difficult to deal with relations between *dan* content paragraphs that are not explicitly indicated using corpus linguistics methods, the analysis in this study is limited to chains of fully expressed connectives at this stage of the research. Among the co-occurrence patterns found in Tables 3 and 4, *shikashi* (however) → *sokode* (therefore), or more generally, ‘adversative connective → additive connective’, has already been examined in detail in Wang (2015a, b), as mentioned above.

From the visualization of potential co-occurrence chains in Figure 1a and Figure 1b, we can further extract longer potential chain patterns that are involved in discourse development.

In the present study, the extraction was restricted to the top 20 PPM and top 20 PMI-valued co-occurring connectives. If all co-occurring examples that meet the co-occurrence condition threshold were included, the number of potential linkage patterns would increase further, but this analysis is left as a future task.

Shikashi (however) and *mata* (also), which occupy a central position in the three corpora in terms of their potential for co-occurrence with other connectives, are also central to several longer co-occurrence chain patterns.

Prominent and possibly formulaic potential connective patterns found in Figure 1a and Figure 1b, based on the 76 different connective co-occurrence pairs[4] in the top 20 examples appearing in Tables 3 and 4 are shown in Table 5 below.

[4] *aruiwa→izureniseyo; uchi→uchi; kokode→mata; kokodewa→mazu; konobaainioite→daiichi; koremade→shikashi; sarani→mata; shikashi→sokode; shikashi→soshite; shikashi→sonotame; shikashi→tsumari; shikashi→mata; shikashi→tatoeba; shikashinagara→sokode; shikashinagara→sonotame; shikashinagara→tokuni; sokode→mata; sokode→gutaitekiniwa; soshite→shikashi; soshite→soshite; soshite→mata; soshite→saigoni; sonota→sonota; soredewa→mazu; daga→toiunomo; tadashi→mata; tsumari→mata; de→de; dewa→dewa; nao→mata; nande→nande; mazu→soshite; mazu→sono-uede; mazu→tsuide; mazu→tsugini; mazu→tsudzuite; mata→konoyōna; mata→konoyōni; mata→sarani; mata→shikashi; mata→shitagatte; mata→sokode; mata→soshite; mata→tsumari; mata→nao; mata→mata; mata→ippō; mata→tatoeba; mochiron→shikashi; mochiron→shikashinagara; mochiron→daga; hitotsuwa→mōhitotsu; hitotsuwa→mōhitotsuwa; hitotsuwa→futatsumewa; hitotsuwa→zenshawa; hitotsumewa→futatsumewa; ippō→mata; ika→konobaainioite; ika→ika; ikadewa→mazu; ikadewa→tsugini; tatoeba→shikashi; tatoeba→mata; saigoni→daiichini; zenshawa→kōshawa; tsugini→saigoni; tashikani→shikashi; daiichi→konobaainioite; daiichi→daiichi; daiichi→dainini; daiichini→dainini; daiichiwa→dainiwa; daisan'ni→daiyon'ni; dainini→daisan'ni; dainiha→daisanwa; kekka→kekka*

Table 5: Prominent potential connective patterns found in Figure 1

	Chain examples	Corpus	Length (No. of connectives)
PPM	*tashikani* (indeed)→***shikashi*** (however)→*tsumari* (namely)	BCCWJ*	3
	tashikani (indeed)→***shikashi*** (however)→*sonotame* (therefore)	HS	3
	tashikani (indeed)→***shikashi*** (however)→*sokode* (therefore)	BCCWJ*, HS	3
	mochiron (of course)→***shikashi*** (however)→*sokode* (therefore)	BCCWJ*, HS	3
	mochiron (of course)→***shikashi*** (however)→***mata*** (again)→*sarani* (again)	BCCWJ*	4
	tashikani (indeed)→***shikashi*** (however)→***mata*** (again)→*sarani* (again)	BCCWJ*	4
	tashikani (indeed)→***shikashi*** (however)→***mata*** (again)→*shitagatte* (therefore)	HS	4
	tashikani (indeed)→***shikashi*** (however)→***mata*** (again)→*sarani* (again)	HS	4
	mochiron ()→***shikashi*** (however)→***mata*** (again)→*ippō* (on the other hand)	BCCWJ*	4
	tashikani (indeed)→***shikashi*** (however)**→*mata*** (again)→*ippō* (on the other hand)	BCCWJ*	4
PMI	*koremade* (hitherto)→***shikashi*** (however)**→***sokode* (therefore)→*gutaitekiniwa* (concretely)	ST	4
	koremade (hitherto)→***shikashinagara*** ()**→***sokode* (therefore)→*gutaitekiniwa* (concretely)	ST	4

If a connective in a chain links two overlapping co-occurring pairs, such as *shikashi* (however) in the chain [*mochiron* (indeed → {*shikashi* (however)} → *tsumari* (namely)], then it is reasonable to assume that a chain consisting of three connectives, linked with the centrally occurring connective can occur in actual discourse.

Table 5 shows the specific possibilities of such potential linkages. The respective ranges of two or more overlapping co-occurrence pairs are indicated by underlining and boldface. The '→' indicates the direction (order) of linkage of the conjunctive expression.

Among the top 20 PMI values, there are highly formulaic 'enumerating' patterns, such as *daiichiwa* (firstly)→*dainiwa* (secondly)→*daisanwa* (thirdly) which are self-explanatory enough and have therefore been omitted from Table 5.

Among the various potential concatenation patterns in Table 5, there are a number of concatenation patterns of co-occurrence pairs, such as the overlapping co-occurrence pairs seen in the top 20 examples of PPM, which are formed around *mochiron* (of course) and *tashikani* (indeed). These patterns are involved with the development of argumentation and are seen in BCCWJ* and Humanities and social sciences data. On the other hand, among the top 20 PMI values, the two concatenation patterns of co-occurrence pairs formed around *shikashi/shikashinagara* (however) → *sokode* (therefore) are only found in the Science and technology data.

In the following Section 5, the potential possibilities of concatenation patterns of co-occurrence pairs presented in this section will be examined in actual discourse, namely in newspaper editorial articles.

5 Verification of co-occurrence chains in actual discourse

This section verifies the potential sequential patterns listed in Table 5 using concrete examples of their use in discourse. For this purpose, various academic monographs and a small corpus of 300 Asahi Shimbun editorials and opinion articles were used. First, we checked the extent to which the 76 co-occurrence pairs of connectives mentioned above overlapped with the chains of connectives found in the examples of actual discourse. Of the 76 co-occurrence pairs, 18 were either used as single co-occurrence pairs or appeared as a part of a chain of multiple connectives. For example, in addition to the examples of the single co-occurrence pairs *shikashi* (however) → *sokode* (therefore), *hitotsuwa* (one) → *mōhitotsuwa* (the other) and *mochiron* (of course) → *daga* (but) , seen in (2)-(4), there are also long chains of co-occurrence pairs in editorials, such as *nanishiro* (anyhow) → *tada* (just) → *shikashi* (however) → *sorewa* (that is) → *mazu* (first) → *soshite* (then) and *shikashi* (however) → *mazu* (first) →*tsugini* (next) → *sonouede* (moreover) → *tatoeba* (for example) → *sonotame* (therefore). Some of the 18 co-occurrence pairs mentioned above, such as *tatoeba* (for example) → *sonotame* (therefore) are included in these chains, and in other examples of long chains. As has been said before, these 18 co-occurrence pairs fulfill the co-occurrence threshold conditions and seem to be used as 'ready-made parts' in the development of discourse.

Some examples of chains containing multiple connectives are given below. The overlapping status of the ready-made co-occurrence pairs in the examples is highlighted by underlining and boldface. The examples extracted from the actual discourse data are shown in the order of the increasing complexity of the chains in which they appear. The first example is a chain formed by just one co-occurrence pair.

(6) 交渉による和平への取り組みが失速すれば、１９８９年のソ連軍撤退の後のような内戦が再燃することになるだろう。
まず米国には、アフガンとパキスタンの双方を説得して協力関係を築いてほしい。
そしてカルザイ大統領は、アフガン人同士で和平をめざす基本に立ち返り、タリバーンとの対話を粘り強く探るべきだ。
（朝日新聞 2011 年 10 月 07 日、朝刊、社説）

Kōshō ni yoru wahei e no torikumi ga shissoku sureba, 1989-nen no Soren-gun tettai no nochi no yōna naisen ga sainen suru koto ni naru darō.
Mazu *Beikoku ni wa, Afugan to Pakisutan no sōhō o settoku shite kyōryoku kankei o kizuite hoshī.*
Soshite *karuzai daitōryō wa, Afugan hito dōshi de wahei o mezasu kihon ni tachikaeri, taribān to no taiwa o nebaridzuyoku sagurubekida.*
(Asahi shinbun 2011 nen 10 tsuki 07 nichi, chōkan, shasetsu)

If negotiated peace efforts stall, civil war will flare up, as it did after the withdrawal of Soviet troops in 1989.
First (*mazu*), the US should persuade both Afghanistan and Pakistan to cooperate with each other.
President Karzai should **then** (*soshite*) return to the fundamentals of peace among Afghans and persistently.
(Asahi Shimbun, 07 Oct 2011, morning edition, editorial)

In example (6), PMI values for the co-occurrence pair *mazu* (first) → *soshite* (then) are BCCWJ* n/a; HS 4.6; ST 5.8.

The argument in (6) is structured as follows. The two reasons supporting the assertion in the first paragraph, i.e., '...civil war will flare up', are introduced by *mazu* (first), which, according to Ishiguro (2008) belongs to the 'organizing-enumerating' type of connectives, and *soshite* (then), which belongs to the 'organizing-coordinating' category. This is a co-occurrence pattern often found in academic papers, but similar patterns such as *mazu* (first) → *tsugini* (next) are also found in the BCCWJ* top 50 PMI examples.

There are also examples of ready-made co-occurrence pairs embedded in longer chains, for example *shikashi* (however) → *sokode* (therefore) → *sonouede* (moreover) in the following chain.

(7) クラウドソーシングのプラットフォーム上では、<…>、発注者は適切な受注者を見つけやすくなっているとする。
しかし、適切に発注者を選んだとしても、<…>やり取りを具体的に見ていく必要があると思われる。
そこで、本章では、クラウドソーシングの発注業務において、発注者と受注者の実際のやり取りを検証し、実際に発注を行った発注者の声を参考にしながら、両者のコミュニケーション上で生まれるやり取りの問題点について明らかにする。
その上で、発注者から提示可能なやり取り文書の改善点を提示することを試みる。
（石黒圭 (編) 2020,14 章）

Kuraudosōshingu no purattofōmu-jōde wa,<…>, hatchū-sha wa tekisetsuna juchū-sha o mitsuke yasuku natte iru to suru.
***Shikashi**, tekisetsu ni hatchū-sha o eranda to shite mo,<…> yaritori o gutaitekini mite iku hitsuyō ga aru to omowa reru.*
***Sokode**, honshōde wa, kuraudosōshingu no hatchū gyōmu ni oite, hatchū-sha to juchū-sha no jissai no yaritori o kenshō shi, jissai ni hatchū o okonatta hatchū-sha no koe o sankō ni shinagara, ryōsha no komyunikēshon-jō de umareru yaritori no mondaiten ni tsuite akiraka ni suru.*
***Sonouede**, hatchū-sha kara teiji kanōna yaritori bunsho no kaizen-ten o teiji suru koto o kokoromiru.*

Suppose that on a crowdsourcing platform, <...>, it is easier for the ordering party to find a suitable order taker.
However (*shikashi*), even if the appropriate ordering party is selected, <...> it is still necessary to look at the specifics of the interaction.
This chapter **therefore** (*sokode*) examines the actual interactions between ordering parties and order takers in the crowdsourcing ordering process, and clarifies the issues that arise in the communication between the two parties, referring to the opinions of the ordering party who actually placed the order.
Moreover (*sonouede*), it attempts to present points for improvement in the exchange documents that can be presented by the ordering party.
(Ishiguro ed. 2020, Chap. 14)

The long chain in example (7) is first broken down into pairs of co-occurring connectives to check the PMI values. Pair 1 *shikashi* (however) → *sokode* (therefore): PMI value: BCCWJ* 2.7; HS 5.2; ST 5.9. Pair 2 *sokode* (therefore) → *sonouede* (moreover): PMI value: no examples meeting threshold conditions.

Example (7) is a good example of a combination of patterned and non-patterned parts in a chain of connectives. In the chain, firstly, in a ready-made co-occurrence pair with a high frequency of *shikashi* (however) → *sokode* (therefore) is used. *Shikashi* (however) introduces the need for verification of the specific interaction between the order taker and the ordering party, and information on how to do this in concrete terms is presented by *sokode* (therefore). Finally, a new, more specific response, added to the first response, is introduced by *sonouede* (moreover): 'to present improvements in the exchange documents that can be presented by the order taker and the ordering party'.

(8) 分析の結果、発注者からみた受注者とのやり取り文書の問題点が明らかになった。以下、**3 点順に挙げていく**。
4.1 作業環境への認識のずれ
まず、発注者と受注者のやり取りの中で、作業環境に対する認識にすれが生まれていた。具体的な例として、A による<…>指示文書は例 1 に、その後の<…>認識のずれを例 2 に示す。
なお、以降、発注者の文言は<…>と示す。
また、発注者の指示文書内の下線部は筆者が記入したものである。
（石黒圭 (編) 2020,14 章）

Bunseki no kekka, hatchūsha kara mita juchūsha to no yaritori bunsho no mondaiten ga akiraka ni natta. Ika, ***3-ten jun ni agete iku.***
4. 1 Sagyō kankyō e no ninshiki no zure <next section title>
***Mazu**, hatchūsha to juchūsha no yaritori no naka de, sagyō kankyō ni taisuru ninshiki ni sure ga umarete ita. Gutaitekina rei to shite, A ni yoru <…> shiji bunsho wa rei 1 ni, sonogo no <…> ninshiki no zure o rei 2 ni shimesu.*
***Nao**, ikō, hatchūsha no mongon wa <…> to shimesu.*
***Mata**, hatchūsha no shiji bunsho-nai no kasenhō wa hissha ga kinyū shita monodearu.*

The analysis reveals problems regarding the exchange of documents with the order taker, as seen the from the point of view of the ordering party. Hereinafter (*ika*) **three points are listed in order**.
4.1 Gaps in the perception of the work environment
First (*mazu*), in the correspondence between the ordering part and the order taker, there was a gap in the perception of the work environment. As a concrete example, the ... instruction document by A is shown in Example 1, and the subsequent gap in ... recognition is shown in Example 2.
Moreover (*nao*), hereafter, the ordering party's wording is shown as ...
In addition (*mata*), the underlined parts in the ordering party's instruction document have been filled in by the author.
(Kei Ishiguro (ed.) 2020, Chap. 14)

First, the long chain of co-occurring connective pairs in example (8), *ika* (hereinafter) → *mazu* (first) → *nao* (moreover) → *mata* (again/in addition) is broken down to check the PMI values.

Pair 1 *ika* (hereinafter) → *mazu* (first): PMI value: BCCWJ* N/A, HS 5.7, ST 4.9. Pair 2 *mazu* (first) → *nao* (moreover): PMI value: no examples meeting threshold conditions. Pair 3 *nao* (moreover) → *mata* (again/in addition): PMI values: BCCWJ* 2.8, HS 4.3, ST 3.8.

(8) is an example of an 'enumerate' chain consisting of two rather weak ready-made co-occurrence pairs. Enumeration is introduced also more specifically following *ika* (hereinafter) by the expression '*3-ten jun ni agete iku* (three points listed in order)', which helps to bridge the section boundary by its explicit cataphoric reference. Pair *ika* (hereinafter) → *mazu* (first) has a moderately high PMI value in the academic data but does not meet the threshold condition in the BCCWJ*. This means that being 'ready-made' is also related to the genre. On the other hand, the last pair *nao* (moreover) → *mata* (again/in addition) meets the threshold condition in all three genres. Both ready-made pairs are connected by the *mazu* (first) → *nao* (moreover), a pair that does not satisfy the threshold condition for co-occurrence in any of the genres examined. This is not surprising since the functions of *mazu* (first), i.e., organizing-enumerating, and *nao* (moreover), i.e., understanding-supplementing, are in conflict.

The next example (9) also contains a long chain of connectives.

(9) 科学的にわかっているのはどこまでか。それをまず明らかにしたうえで、安心して暮らせる環境を取り戻すための**放射線対策を提案している**。

まず、最も低いレベルの放射線の影響として科学的に立証され、国際的に認められているのは、１００ミリシーベルト浴びるとがんで亡くなるリスクが 0.５％高まる、というものだ。これより低いと、健康への影響は科学的にはわからない。**しかし**、国際放射線防護委員会（ＩＣＲＰ）は健康を守る立場から、線量に比例してがんのリスクが高まると仮定して防護策をとり、線量を減らしていくことを求めている。日本もこれに従っている。

また、同じ量でも短期間に浴びた方がリスクは大きく、また同じ量なら、外部被曝（ひばく）も内部被曝もその影響は同じだと、国際的に認められている。＜…＞

除染には優先順位をつけ、目標を立てて段階的に進めるよう求めた。

そして、子どもの健康に不安を感じる人が多いことを考え、子どもがいる環境の除染を優先すべきだとした。

（朝日新聞 2011 年 12 月 17 日、朝刊、社説）

Kagakutekini wakatte iru no wa doko made ka. Sore o mazu akiraka ni shita ue de, anshin shite kuraseru kankyō o torimodosu tame no ***hōshasen taisaku o teian shite iru.***
Mazu*, mottomo hikui reberu no hōshasen no eikyō to shite kagakutekini risshō sa re, kokusai-teki ni mitome rarete iru no wa, 100 mirishīberuto abiru to gan de nakunaru risuku ga 0. 5-Pāsento takamaru, to iu monoda. Kore yori hikui to, kenkō e no eikyō wa kagakutekini wa wakaranai.*
Shikashi*, kokusai hōshasen bōgo iinkai (ICRP) wa kenkō o mamoru tachiba kara, senryō ni hirei shite gan no risuku ga takamaru to katei shite bōgo-saku o tori, senryō o herashite iku koto o motomete iru. Nihon mo kore ni shitagatte iru.*
Mata, *onaji ryō demo tankikan ni abita kata ga risuku wa ōkiku, mata onaji ryōnara, gaibuhibaku (hibaku) mo naibu hibaku mo sono eikyō wa onajida to, kokusai-teki ni mitome rarete iru. <…> Jo some ni wa yūsen jun'i o tsuke, mokuhyō o tatete dankai-teki ni susumeru yō motometa.*
Soshite*, kodomonokenkō ni fuan o kanjiru hito ga ōi koto o kangae, kodomo ga iru kankyō no jo some o yūsen subekida to shita.*
(Asahi shinbun 2011 nen 12 tsuki 17 nichi, chōkan, shasetsu)

How much is known about this scientifically? This is to be clarified first and then **radiation countermeasures are proposed** to restore a safe environment for people to live in.
First (*mazu*), the effect of the lowest level of radiation that is scientifically proven and internationally accepted is that exposure to 100 millisieverts increases the risk of dying from cancer by 0.5%. At lower levels, the effects on health are not scientifically proven. **However** (*shikashi*), the International Commission on Radiological Protection (ICRP) assumes from the standpoint of protecting health that the risk of cancer increases in proportion to the dose and calls for protective measures to be taken and doses to be reduced. Japan follows this approach.
Again (*mata*), it is internationally recognized that the risk is greater if the same amount of radiation is received over a short period of time, and that the effects of external and internal exposure are the same if the same amount is received. <...>
The decontamination process should be prioritized, with goals set up and carried out step by step.
Then (*soshite*), he stated that considering that many people are concerned about the health of children, priority should be given to the decontamination of environments where children are present.
(Asahi Shimbun, 17 Dec 2011, morning edition, editorial)

The long chain of connectives in (9), *mazu* (first) → *shikashi* (however) → *mata* (again) → *soshite* (then), is introduced by a cataphoric reference in the immediately preceding paragraph, i.e., '*hōshasen taisaku o teian shite iru* (radiation countermeasures are proposed)'.

Considered from the point of view of the text organization, this chain actually consists of only three directly interacting connectives. The pair *mazu* (first) → *shikashi* (however) is functioning only in the local *dan* content paragraph and based on its PMI value, the pair does not satisfy the threshold condition. It is therefore an ad hoc co-occurrence of connectives.

The actual chain is thus *mazu* (first) → *mata* (again) → *soshite* (then). This chain has first to be broken down into co-occurring pairs so that the PMI values can be checked.

Pair 1 *mazu* (first) → *mata* (again) PMI values: BCCWJ* 1.2, HS 3.8, ST 3.7 (In BCCWJ* the pair does not meet the threshold condition and is therefore not considered a co-occurring pair in this genre). Pair 2 *mazu* (first) → *mata* (again) PMI value: HS 4.2. The pair does not satisfy the threshold condition in BCCWJ* and ST. Pair 3 *mata* (again) → *soshite* (then) PMI value: 3.6 in HS. The pair does not satisfy the threshold condition in BCCWJ* and ST.

In example (9), in order to specify the 'proposed radiation measures' by the Government, the text is basically organized as a 'triple jump' of segments, introduced by *mazu* (first), *mata* (again), and finally *soshite* (then), all connectives being of the 'organize-enumerate' type.

In the *dan* content paragraph introduced by *mazu* (first), a discussion leading to a standard limit on radiation doses is presented. In contrast to the 'national standards', the 'international standards' are introduced locally, within the same *dan* content paragraph, by *shikashi* (however). Therefore, *shikashi* (however) does not form a content paragraph-based co-occurring pair with *mazu* (first) and is therefore not a part of the rest of the chain. It is therefore *mata* (again) of 'organize-coordinate' type that can be regarded as co-occurring with *mazu* (first). *Mata* (again) introduces the *dan* content paragraph about a link between the radiation dose and the time of exposure and further also a specific measure based on that link.

Finally, in contrast to the *dan* content paragraph introduced by *mata* (again), *soshite* (then) of the 'organize-coordinate' type introduces the last *dan* content paragraph which deals with the decontamination of the environment in which the children are located.

Based on the relatively high PMI values, here the chain *mazu* (first) → *mata* (again) → *soshite* (then) is formed by the overlapping *mazu* (first) → *mata* (again) and *mata* (again) → *soshite* (then), both of which are 'ready-made' co-occurrence pairs. So, this chain can be regarded as being formed directly by 'ready-made' co-occurrence pairs.

In the two-item co-occurrence pairs in examples (2), (3), (4) and (6) seen in the previous section and above, the PMI values of the co-occurrence criteria are at least 3.5 at the lowest, almost twice as big as the co-occurrence threshold condition. This means that these co-occurrence pairs are often used as ready-made elements. They may be used not only as single pairs but also as a part of longer chains. Examples (7) and (8) are cases where the chain contains one or two ready-made pairs. Example (9), on the other hand, is an example where the ready-made pairs *mazu* (first) → *mata* (again) and *mata* (again) → *soshite* (then) overlap over *mata* (again). Again, ready-made co-occurrence pairs are used to develop the discourse. Based on PMI values involved in all these patterns, we can consider that the resulting overlapping pattern with three connectives, i.e., *mazu* (first) → *mata* (again) → *soshite* (then), and other co-occurrence chains formed with high PMI values are also 'ready-made' co-occurrence patterns.

Needless to say, from the speaker/writer's point of view 'ready-made' patterns are useful for discourse development because they reduce the discourse planning load. Because of their formulaicity, they also contribute to the predictability of discourse development from the listener's/reader's point of view. In other words, they reduce the cognitive load in both production and processing, thus contributing to the fluency of linguistic exchange.

The above observations thus point out to the existence of 'ready-made' co-occurrence chains that are longer than 'ready-made' co-occurrence pairs and they also clarify the overall role such chains play in discourse. Based on this, the answers to RQ2 and RQ3 can also be considered affirmative.

6 Discussion and conclusions

The distant co-occurrence of connectives has received increasing attention over the last ten or fifteen years. The present study is an exploratory study, aimed at determining the presence or absence of two-item distant co-occurrence patterns (RQ1), as well as the presence or absence of multiple-item distant co-occurrence patterns (RQ2), and the role of these patterns in discourse, especially in relation to the cognitive load needed to process the incoming discourse (RQ3). In order to identify potentially formulaic co-occurrences, we used general written material (the BCCWJ* corpus) and academic paper material (the Humanities and Social Sciences papers corpus and the Science and Technology papers corpus). The conditions for distant co-occurrence were somewhat more relaxed than in traditional collocation studies, with a co-occurrence frequency > 10, PMI value > 2, and Dice coefficient > 0.01. As for the co-occurrence cases meeting these conditions, 87 were found in BCCWJ*, 181 in Science and Technology paper data, and 202 in Humanities and Social Sciences papers data.

In order to identify reliable co-occurrences, in the present study, only the top 20 examples from each corpus in terms of PPM and PMI values were included in the analysis.

In terms of co-occurrences with high PMI values, the BCCWJ* data showed a low correlation with the PPM value, while the same correlation was relatively high in both academic corpora. This suggests that the range of combinations of co-occurring items in each of the academic corpora is narrower than in the general data represented in BCCWJ* and that the degree of formulaicity is consequently higher.

The vast majority of co-occurrence pairs that meet the aforementioned co-occurrence threshold conditions in the three corpora can intuitively be regarded as 'ready-made' co-occurrence pairs. Among the top 20 PMI values, typical co-occurrence patterns of connectives such as *hitotsuwa* (one)→ *mōhitotsuwa* (the other), both belonging to the 'organize-enumerating' type especially are prominent.

In addition to the top 20 PMI cases examined here, the majority of other co-occurrence cases that meet the co-occurrence threshold conditions also appear to be valid as two-item co-occurrence 'ready-made' pairs. The examination of the top 20 PPM cases revealed a more diverse pattern: in addition to many similarities with the top 20 PMI values, genre-specific differences were also noticeable. The answer to RQ1 is therefore in the affirmative. There is a need to investigate these differences in more detail in the future, taking into account, for example, teaching Japanese as a second language.

The multiple connective co-occurrences, i.e., chains of co-occurrences, were then identified based on the visualization of co-occurrence pairs by means of directed graphs. The visualization revealed similarities and differences between the top 20 PPM cases and the top 20 PMI cases. The similarities between genres, i.e., general vs. academic, are more pronounced in the top 20 PPM co-occurrences. In particular, *shikashi* (however) and *mata* (again) were found to be two centers around which a large number of co-occurrences of connectives are formed. At the same time, many other combinations of connectives were also present.

On the other hand, the top 20 PMI values are dominated by examples of longer chains of 'organize-enumerate' type of connective co-occurrences. As with the BCCWJ* data, there are few other types of co-occurrences, and the potential for combination with longer chains of connectives seems to be limited. In contrast, in both academic corpora, potential opportunities for longer chain formation other than of 'organize-enumerate' type, were revealed.

The chains of connectives visualized in the directed graphs in Figure 1a and Figure 1b, created on the basis of the co-occurrence data are to be understood as potential chains that can be used in the actual development of discourse, as also shown in Example (8) in Section 5. More specifically, they can be seen as 'ready-made' patterns that can potentially be used in argumentative prose such as academic papers and editorials.

Next, these ‘ready-made’ patterns, belonging to the realm of the possible, have been examined in actual discourse data. Specifically, instead of three corpora, a small corpus of editorial and opinion articles from the Asahi Shimbun (300 articles) and a specialist humanities monograph were used to examine chains containing multiple co-occurring connectives.

The extraction of multiple connective co-occurrence chains from the Asahi Shimbun data yielded 18 co-occurrence pairs that were included in the set of 76 co-occurrence pairs extracted based on the top 20 PPM and PMI values. Some of these 18 co-occurrence pairs here were frequently found to partially overlap in these extracted chains. This suggests the existence of longer systematic co-occurrence chains and also sheds light on the strategies for forming longer chains. In other words, the co-occurrence pairs of connectives as ‘ready-made’ parts play an important role in the formation of longer chains. The answer to RQ2 is therefore also affirmative.

The interpretation and positioning of such chains of systematically co-occurring connectives are in some ways similar to what Wray (2002) and others refer to as formulaic expressions, they can be seen as a kind of formulaicity and therefore contribute to discourse development. However, they differ from the conventional notion of formulaicity in that they are observed at the discourse level. On the other hand, the scope of formulaic expressions treated in conventional studies, such as Wray (2002) and Tanaka (2016), is limited to a single sentence. Therefore, many interesting findings from the conventional research on formulaic expressions cannot be directly applied to the distant co-occurrence of connectives.

To put the regularities observed in distant co-occurrence of connectives into proper perspective, Ishiguro’s (2008) view of ‘strategic usage’ patterns is one reasonable way of looking at the phenomenon. At the same time, Bourdieu’s notion of *habitus* also seems to provide a valid framework for its interpretation (see Bourdieu, 1991, 1994). In Bourdieu’s terms, the ‘ready-made’ chains of systematically co-occurring connectives in discourse reflect argumentative patterns internalized by the writer/speaker in the course of linguistic activity. Therefore, while the co-occurrence chain patterns observed here are part of the individual *habitus*, they can also be interpreted as forming part of the collective *habitus* of a particular linguistic community, since listeners/readers also internalize these patterns. Peers in an academic discipline, for example, are a good example of this. Such internalized ‘ready-made’ patterns contribute to the ease of organization and development of discourse on the part of the author and to the ease of comprehension on the part of the reader/listener. On the other hand, there is a negative aspect to this phenomenon. Namely, by directing the flow of thought into the predictable habitual channels, these ready-made patterns of argumentation can hinder the conception of new ideas and understanding.

As for the diversity of the usage of connectives, the greater diversity is found in the general BCCWJ* data as compared to the academic data. The reason is the higher need for accuracy in the transmission of academic data, as compared to general use. This is also one of the conceivable motives behind the more pronounced formulaicity in academic communication.

In conclusion, the aim of the present study was very limited: to test Ishiguro's (2008) predictions about the 'strategic usage' patterning of connectives and to ascertain the potential contribution of such patterning to discourse development. A tentative conclusion can be drawn that 'strategic usage' patterning is indeed a widely recognized systematic phenomenon that contributes to discourse development and understanding. The present study has also shown that directed graph visualization has a good potential for identifying such 'strategic usage' patterning. These preliminary result needs to be further tested and elaborated, both quantitatively and qualitatively, using additional linguistic material. The findings are expected to have applications in language teaching, particularly academic writing and teaching Japanese as a second language, in critical discourse analysis, and in language theory in general.

Acknowledgements

We are grateful to Professor Kei Ishiguro for his intellectual stimulation over the past few years. We are also grateful to Jinbo Wang for stimulating our interest in this field and for sharing some of her research materials. We would also like to thank Professor Yuriko Sunagawa and Professor Nagisa Moritoki for their useful comments. Last but not least, we would like to express our sincere gratitude to the members of the '*Setsuzoku-ken*' (connectives research) discussion group, who continue to share our interest in Japanese connectives. Needless to say, the authors are responsible for all the shortcomings of this study.

References

Abekawa, T. 阿辺川武, Nishina, K. 仁科喜久子, Yagi, Y. 八木豊, & Hodošček, B. ホドシチェック・ボル (2020). Nihongo setsuzoku hyōgen no keiryō-teki bunseki ni motodzuku shidō-hō no teian 「日本語接続表現の計量的分析に基づく指導法の提案」[Proposal for a teaching method based on a quantitative analysis of Japanese connectives], *Keiryōkokugogaku, 『計量国語学』, 32*(7), 387-401.

Bekeš, A. (2008). *Text and Boundary: A Sideways Glance at Textual Phenomena in Japanese*. Ljubljana: Ljubljana University Press.

Bekeš, A. (2012). Suppositional Adverb-based Brackets in Discourse. In R. Tomiya 富谷玲子 & M. Tsutsumi 堤正典 (Eds.), *Modariti to gengokyouiku 『モダリティと言語教育』 [Modality and language education]* (pp. 21-37). Tokyo: Hituzi Shobo ひつじ書房.

Bourdieu, P. (1991). *Language and symbolic power* (J. B. Thompson, Ed., G. Raymond & M. Adamson, Trans.). Cambridge: Polity Press.

Bourdieu, P. (1994). *Raisons pratiques: sur la théorie de l'action*. Paris: Seuil.

de Beaugrande, R., & Dressler, W. U. (1981). *Introduction to text linguistics*. London: Longman.

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory Social Network Analysis with Pajek*. Cambridge: Cambridge University Press.

de Saussure, F. (1966). *Cours de linguistique générale* (W. Baskin, Trans.) *Course in General Linguistics*, New York: Mac Graw-Hill. (Original work published 1916)

Ichikawa, T. 市川孝 (1978). *Kokugo kyōiku no tame no bunshō-ron gaisetsu 『国語教育のための文章論概説』[Introduction to text theory for teaching of Japanese as the first language]*. Tokyo: Kyoiku Shuppan 教育出版.

Ishiguro, K. 石黒圭 (2008). *Bunshō wa setsuzoku hyōgen de kimaru 『文章は接続表現で決まる』 [Sentence is determined by the connective expressions]*. Tokyo: Kobunsha 光文社.

Kaneyasu, M., Ajioka, M., Kawanishi, Y., & Iwasaki, S. (2015). Mikan yo Mikan: Formulaic Constructions and Their Implicature in Conversation. *Japanese/Korean Linguistics 21*, 199-213. Stanford, CA: CSLI Publications.

Kolesnikova, O. (2016). Survey of Word Co-occurrence Measures for Collocation Detection. *Computación y Sistemas, 20*(3), 327-344. doi: 10.13053/CyS-20-3-2456

Kudo, H. 工藤浩 (2000). Fukushi to bun no chinjutsu-teki taipu 「副詞と文の陳述的タイプ」 [Adverbs and the type of medus in sentence]. In Y. Nita 仁田義雄 & T. Masuoka 益岡隆志 (Eds.), *Nihongo no bunpō 3 — modariti 『日本語の文法３—モダリティ』[Grammar of Japanese 3: Modality]*. Tokyo: Iwanami Shoten 岩波書店.

Minami, F. 南不二雄 (1974). *Gendai nihongo no kōzō 『現代日本語の構造』 [The Structure of modern Japanese language]*. Tokyo: Taishukan Shoten 大修館書店.

Minami, F. 南不二雄 (1993), *Gendai nihongo bunpō no rinkaku 『現代日本語文法の輪郭』 [Theoutline of modern Japanese grammar]*. Tokyo: Taishukan Shoten 大修館書店.

Mrvar, A., & Batagelj, V. (2022). *Pajek Programs for Analysis and Visualization of Very Large Networks - Reference Manual ver. 5.16* <http:// mrvar.fdv.uni-lj.si/pajek/> last accessed June 1, 2023.

Noda, H. 野田尚史 (1995). Bun no kaisō koozō kara mita shudai to toritate 「文の階層構造と主題の取り立て」[Theme and extrapolation viewed from the hierarchical sentence structure]. In T. Masuoka 益岡隆志 et al. (Eds.), *Nihongo no shudai to toritate 日本語の主題と取り立て [Theme and extrapolation in Japanese]* (pp. 1-35). Tokyo: Kurosio Publishers.

Petrovic, S., Snajder J., Dalbelo Basic B., & Kolar, M. (2006). Comparison of Collocation Extraction Measures for Document Indexing. *Journal of Computing and Information Technology, 14*(4), 321-327. (doi:10.2498/cit.2006.04.08).

Sakuma, M. 佐久間まゆみ (2012). Bunshō danwa no bunseki tan'i 「文章・談話の分析単位」 [Units of analysis intext and discourse]. *Gengo - serekushon『言語』セレクション 1,* 93-100.

Sakuma, M. (2019). Units for the analysis of Japanese written text and spoken discourse. In I. Srdanović & A. Bekeš (Eds.), *The Japanese Language from an Empirical Perspective:*

Corpus-based studies and studies on discourse (pp. 11-30). Ljubljana: University of Ljubljana Press.

Srdanović Erjavec, I., Bekeš, A., & Nishina, K. (2007). Cluster analysis of suppositional adverbs and clause-final modality. *Asian and African Studies, 11*(3), 21-31.

Srdanović, I., Hodošček, B., Bekeš. A, & Nishina, K. (2009). Uebukōpasu to kensaku shisutemu o riyō shita suiryō fukushi to modariti keishiki no enkaku kyōki chūshutsu to nihongo kyōiku e no ōyō 「ウェブコーパスと検索システムを利用した推量副詞とモダリティ形式の遠隔共起抽出と日本語教育への応用」 [Distant co-occurrence extraction of inferred adverbs and modality forms using a web corpus and search system and its application to Japanese language teaching]. *Keiryōkokugogaku 『自然言語処理』, 16*(4), 29-46.

Tanaka, S. 田中茂範 (2016). Dainigengo hattatsu ni okeru kan'yō hyōgen-ryoku 「第二言語発達における慣用表現力」 [Conventional expressivity in second language development]. *ARCLE Review 10*, 40-52. <https://www.arcle.jp ' research ' books ' data ' html ' data ' pdf ' vol10_4-1.pdf>: last accessed January 15, 2023.

Wang, J. 王金博 (2015a). Ronsetsu bun ni okeru setsuzoku hyōgen no `enkaku kyōki' ni tsuite no kenkyū: Shinbun shasetsu no `shikashi' to `sokode' o chūshin ni 『論説文における接続表現の「遠隔共起」についての研究：新聞社説の「しかし」と「そこで」を中心に』 [A study on 'distant co-occurrence' of connectives in editorial writing: focusing on 'shikashi (but)' and 'sokode (there)' in newspaper editorials]. PhD dissertation <https://tsukuba.repo.nii.ac.jp/record/37004/files/DA07517.pdf>, last accessed January 20, 2023

Wang, J. 王金博 (2015b). Ronsetsu bun no bunmyaku tenkai ni okeru setsuzoku hyōgen `shikashi' to `sokode' no enkaku kyōki' 「論説文の文脈展開における接続表現「しかし」と「そこで」の遠隔共起」[Distant co-occurrence of the connectives 'shikashi (but)' and 'sokode (there)' in the contextual development of editorial texts]. *Kokusai Nihon kenkyū 『国際異本研究』 7*, 97-110. <https://japan.tsukuba.ac.jp/content/uploads/sites/43/2022/02/JIAJS_Vol7_PRINT_06_Wang.pdf>, last accessed January 20, 2023.

Wray, A. (2002). *Formulaic Language and the Lexicon*. New York: Cambridge University Press.

Wray, A. (2017). Formulaic Sequences as a Regulatory Mechanism for Cognitive Perturbations During the Achievement of Social Goals. *Topics in Cognitive Science*, *9*(3), 569-587.

Additional analyzed materials

Asahi Shimbun: 300 Editorials and Opinion articles (31 Jul 2011 - 31 Dec 2011). 朝日新聞社説・意見 300 記事（2011 年 07 月 31 日～2011 年 12 月 31 日）

Ishiguro, K. 石黒圭 (Ed.) (2020). *Bijinesu bunsho no ōyōgengo-gaku-teki kenkyū — kuraudosōshingu o mochiita bijinesu nihongo no takaku-teki bunseki 『ビジネス文書の応用言語学的研究―クラウドソーシングを用いたビジネス日本語の多角的分析』ひつじ書房 [Applied linguistic research of business documents: a multidimensional analysis of business Japanese using crowdsourcing data]*. Tokyo: Hitsuji Shobo.

Directed graph application

Pajek: analysis and visualization of very large networks <http://mrvar.fdv.uni-lj.si/pajek/>, last accessed June 30, 2023.