Journal of Quantitative Linguistics
2004, Vol. 11, No. 3, pp. 233–250
DOI: 10.1080/09296170420003l4976

Routledge
Taylor & Francis Group

Entropy and Redundancy of Japanese Lexical and Syntactic Compound Verbs

Katsuo Tamaoka[1], Hyunjung Lim, and Sakai Hiromu
Hiroshima University, Japan

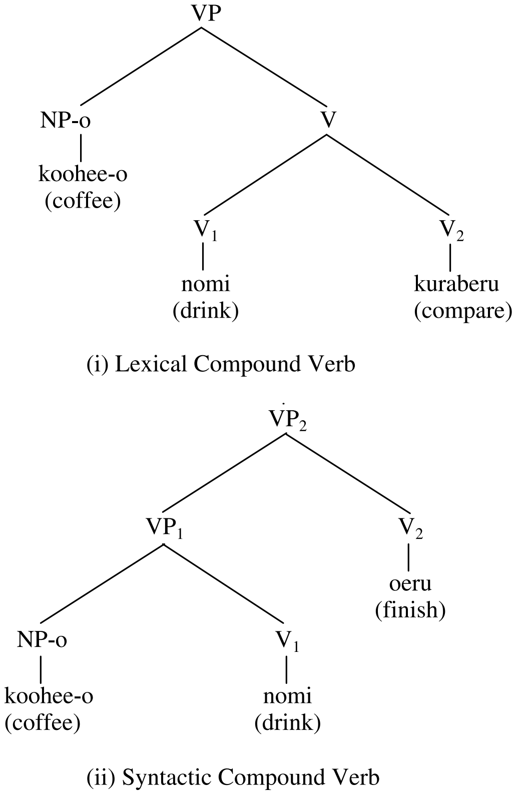
ABSTRACT

The present study investigated Japanese lexical and syntactic compound verbs ($V_1 + V_2$) using Shannon's concept of entropy and redundancy calculated using corpora from the *Mainichi Newspaper* and a collection of selected novels. Comparing combinations of a V_2 verb with various V_1 verbs, syntactic compounds were higher in entropy than lexical ones while neither differed in redundancy. This result suggests that V_2 verbs of syntactic compounds are likely to combine with a wider range of V_1 verbs than those of lexical compounds. Two exceptional V_2 verbs, *komu* and *ageru*, both of which create lexical compounds, showed a wide variety of combinations with V_1 and therefore act like prefixes in English. Comparing V_2 verbs in the two corpora, the V_2 *eru*, which adds the meaning of "possibility" to a V_1, functions like the auxiliary verb "can" in English and seems to be a favored expression in newspapers. In contrast, the V_2 *komu*, adds the meaning of "internal movement" similar to the preposition "into" in English and appears to be preferred in the novels to enrich the expression of lexical compounds. In general, both lexical and syntactic compounds were used similarly in both corpora.

1. INTRODUCTION

The Japanese language frequently combines two verbs to make compound verbs. These compounds are further classified into two groups, lexical and syntactic compounds (Kageyama, 1993, 1999a, 1999b). Lexical compound verbs are mostly limited to lexically-specified combinations of the firstly-positioned verbs (V_1) and the secondly-positioned verbs (V_2) to construct conventional meanings such as *tabe* + *kuraberu* (eat + compare), *nomi* + *aruku* (drink + walk) and *kaki* + *toru* (write + take). For the syntactic

[1] Address correspondence to: Katsuo Tamaoka, International Student Center, Hiroshima University, 1-1, 1-chome, Kagamiyama, Higashi-Hiroshima, 739-8524 Japan. Tel.: 0824-24-6288. Fax: 0824-24-6288. E-mail: ktamaoka@hiroshima-u.ac.jp



(i) Lexical Compound Verb

(ii) Syntactic Compound Verb

Fig. 1. Syntactic structure of lexical and syntactic compound verbs.
Note: NP-o refers to an accusative case-marked noun phrase.

structure of lexical compound verbs such as those shown in Figure 1, two verbs of V_1 *nomi* (drink) and V_2 *kuraberu* (compare) construct a single compound verb *nomikuraberu*, and further construct a verb phrase with an accusative noun phrase *koohee-o* (coffee). In contrast, syntactic compound verbs generally do not have idiosyncratic lexical combinations of two verbs, so that they are rather semantically transparent, such as *tabe* + *hazimeru* (eat + begin), *nomi* + *oeru* (drink + finish) and *kaki* + *naosu* (write + fix). For the syntactic compound verbs, the accusative noun phrase *koohee-o* and V_1 *nomi* construct the first verb phrase VP_1, and are further combined with the V_2 *oeru* (finish) to produce the second verb phrase VP_2. These differences in syntactic structure predict frequencies of appearance; certain verb combinations are seen with a high frequency for lexical compound verbs, whereas a variety of combinations based upon V_2 are seen for syntactic compound verbs. The present study counted token and type frequencies of compound verbs based on V_2, using two different corpora, the *Mainichi Newspaper* and the older novel collection of *Aozora Bunko.* Furthermore, using these frequency figures, mathematical indexes of ‘‘entropy’’ and ‘‘redundancy’’ were calculated in order to clarify the differences between lexical and syntactic compound verbs.

2. ENTROPY AND REDUNDANCY

The concepts of entropy and redundancy were first developed by an American mathematician, Claude Elwood Shannon (1916–2001), in his well-known work, *A Mathematical Theory of Communication* (1948). Since these concepts can be applied to a wider range of corpus size, characteristics of lexical and syntactic compound verbs are directly compared in appearance in both corpora of the *Mainichi Newspaper* and *Aozora Bunko.*

Entropy is an index for the degree of disorder or chaos (for details, see Hori, 1979; Kaiho, 1989). For compound verbs, entropy refers to how randomly V_2 is combined to various V_1, and it is calculated using the following formula:

$$H = \sum_{j=1}^{j} p_j \log_2 p_j$$

In the present study, the entropy of compound verbs was calculated on the basis of the second verb V_2. For example, V_2 *aruku* (walk), which produces lexical compound verbs, appeared to be combined with 18 different verbs in the newspaper corpus. The total of V_2 appearance with other verbs was 44, with *uri* + *aruku* (a combination of sell and walk) having the highest frequency at 7, and *tazune* + *aruku* (a combination of visit and walk) was the second highest at 6. The p in the formula stands for the probability of appearance for a specific compound verb among all the compounds created with V_2. In the case of V_2 *aruku*, p is 0.159, as calculated by dividing 7 by 44. The formula $\log_2 P_j$ for *aruku* is simply counted as $\log_2 0.159 = -2.652$. Then, $p_j \log_2 p_j$ for the V_2 *aruku* is -0.422 (the result of 0.159×-2.652). In the same manner, the values for the remaining 17 compound verbs were also calculated.

The entropy of the V_2 *aruku* is finally determined as 3.780 by adding all these scores and dividing by -1. The calculation for *aruku* is as follows:

$$\begin{aligned} H = &-(7/44)\log_2(7/44) - (6/44)\log_2(6/44) - (5/44)\log_2(5/44) \\ &- (5/44)\log_2(5/44) - (3/44)\log_2(3/44) - (3/44)\log_2(3/44) \\ &- (2/44)\log_2(2/44) - (2/44)\log_2(2/44) - (2/44)\log_2(2/44) \\ &- (1/44)\log_2(1/44) - (1/44)\log_2(1/44) - (1/44)\log_2(1/44) \\ &- (1/44)\log_2(1/44) - (1/44)\log_2(1/44) - (1/44)\log_2(1/44) \\ &- (1/44)\log_2(1/44) - (1/44)\log_2(1/44) - (1/44)\log_2(1/44) = 3.780 \end{aligned}$$

In addition to entropy, another well-known mathematical concept by Shannon is redundancy, which refers to the degree of superfluousness. For compound verbs, it implies frequency bias in appearance, indicating, for instance, repeatedly used expressions in the same corpus. Redundancy is determined using the following formula.

$$R = (1 - H/H_{\max}) \times 100(\%)$$

H refers to entropy, whereas $H_{\max}$ indicates maximum entropy. Maximum entropy for compound verbs implies that any verb would be combined with V_2 in the same probability, as produced by the following formula.

$$H_{\max} = \log_2 J$$

As with the aforementioned example of *aruku*, $H_{\max}$ is calculated to be 4.170 using logarithm $\log_2$ of type frequency J. This figure shows the entropy for 18 different verbs (V_1) equally (i.e., the same token frequency) when combined with V_2 *aruku*. Redundancy for *aruku* is now calculated as a percentage: the result of the figure for entropy (3.780) multiplied by the maximum entropy (4.170) is subtracted from 1.

At a glance, redundancy seems to have a negative correlation of -1 with entropy, but the entropy and redundancy for V_2 correlated moderately as -0.5489 for the newspaper corpus ($n = 48$) and -0.217 for the novel corpus ($n = 37$). Therefore, these two concepts can be treated as different measurements. An advantage of entropy and redundancy is that they can be used with different sizes of relatively large corpora (excluding smaller sizes), so the present study utilized these concepts to compare characteristics of lexical and syntactic compound verbs in both the ‘‘newspaper and novels’’ corpora. As shown in Figure 1, V_2 of syntactic compound verbs is freely combined with VP_1 whereas V_2 of lexical compound verbs appears as a part of a $V_1 + V_2$ single verb. Therefore, it is hypothesized that syntactic compound verbs would be more likely to be higher in entropy and lower in redundancy than lexical compound verbs.

3. FREQUENCY, ENTROPY AND REDUNDANCY OF COMPOUND VERBS FROM THE *MAINICHI NEWSPAPER*

The present study used editions of the *Mainichi Newspaper* published from 1991 to 1994, consisting of a total token frequency of 88,454,573 words. A total of 48 V_2 verbs were selected from the pool of 88 lexical compound and 21 syntactic compound verbs (108 V_2 candidates), using more than 10 token frequencies out of 88,454,573 words appearing in the *Maichini Newspaper* as a baseline.

3.1. Comparison of Lexical and Syntactic Compound Verbs

One-way analyses of variance (ANOVAs) were conducted with lexical and syntactic compound verbs. The ANOVA for entropy indicated that lexical compound verbs ($n = 37$, M = 2.97, SD = 1.16) were significantly lower than syntactic compound verbs ($n = 11$, M = 4.38, SD = 1.95) [$F(1, 46) = 8.95$, $p < .01$]. In contrast, the ANOVA for redundancy indicated no significant difference between lexical compound verbs ($n = 37$, M = 18.43, SD = 15.72) and syntactic compound verbs ($n = 11$, M = 24.96, SD = 25.88) [$F(1, 46) = 0.31$, *n.s.*]. These results suggest that V_2 verbs of syntactic compounds combined irregularly with various V_1 verbs in comparison to the V_2 verbs of the lexical compounds. As depicted in Figure 1, the two verbs of V_1 and V_2 for the lexical compounds strongly combine as lexically idiosyncratic, seeming to appear as $V_1 + V_2$ on a regular basis. In contrast, since V_2 verbs of syntactic compounds can be combined with a variety of V_1 verbs, they are likely to appear irregularly in various $V_1 + V_2$ combinations. However, both the lexical and syntactic compound verbs were equally redundant; this is probably because both include a variety of compounds appearing only once or a few times.

3.2. Classification of Lexical and Syntactic Compound Verbs in the Newspaper Corpus

In order to examine V_2 individual differences in the newspaper corpus further, all 48 compound verbs were plotted on the basis of their entropies and

Table 1. Frequency, Entropy and Redundancy of Lexical and Syntactic $V_1 + V_2$ Compound Verbs in the *Mainichi Newspaper*.

#	Type	V_2 verbs		V_2 token frequency	V_1 total token	V_1 type frequency	V_1 & V_2 token frequency	Entropy	Redundancy (%)
		Japanese	Phonetic						
1	Lexical	込む	komu	295	1,098,690	81	278	5.76	9.10
2		あげる	ageru	2,914	45,880	57	174	5.30	9.20
3		切れる	kireru	543	64,292	44	119	4.66	14.73
4		取る	toru	5,947	53,493	33	94	4.39	13.04
5		回る	mawaru	1,021	17,989	27	61	4.27	10.12
6		つく	tuku	2,354	8,906	19	45	3.81	10.34
7		歩く	aruku	1,554	30,414	18	44	3.78	9.35
8		上がる	agaru	1,808	40,283	31	229	3.69	25.56
9		継ぐ	tugu	355	20,382	15	33	3.68	5.88
10		死ぬ	sinu	1,376	16,929	13	14	3.66	0.97
11		たてる	tateru	632	434,024	16	55	3.66	8.46
12		かかる	kakaru	4,764	61,171	14	25	3.62	4.83
13		替える	kaeru	135	14,613	15	40	3.58	8.44
14		いれる	ireru	2,114	13,410	13	19	3.58	3.35
15		刺す	sasu	333	21,434	11	12	3.42	1.19
16		返す	kaesu	609	9,679	23	45	3.36	25.72
17		出る	deru	7,153	8,513	18	56	3.21	22.94
18		こめる	komeru	110	27,164	12	23	3.13	12.63
19		落ちる	otyiru	755	12,213	11	33	3.07	11.26
20		落とす	otosu	599	4,939	11	20	3.05	11.94
21		おろす	orosu	253	7,527	10	58	2.88	13.29
22		きる	kiru	1,529	136,575	60	496	2.77	53.04
23		入る	iru	6,425	19,260	10	25	2.76	17.06
24		飛ばす	tobasu	244	1,056	7	13	2.57	8.62
25		つける	tukeru	460	20,528	9	34	2.51	20.80
26		倒す	taosu	140	805	7	14	2.41	14.02
27		殺す	korosu	444	1,614	6	14	2.35	8.98

28		起こす	okoru	1,392	3,506	6	19	2.07	19.76
29		渡る	wataru	573	4,639	7	41	2.00	28.81
30		おりる	oriru	431	10,443	7	27	1.68	40.21
31		のぼる	noboru	3,417	13,187	4	11	1.68	16.16
32		返る	kaesu	91	1,487	5	23	1.61	30.68
33		広げる	hirogeru	856	7,812	4	12	1.42	29.09
34		渡す	watasu	491	48,068	3	10	1.36	14.13
35		くだる	kudaru	124	14,816	3	15	1.27	19.69
36		知る	siru	1,993	2,830	5	24	1.14	50.96
37		合わせる	awaseru	1,109	38,142	15	61	0.88	77.38
Means				1,496	63,154	17.57	62.59	2.97	18.43
Standard Deviation				1,804	186,738	17.14	92.97	1.14	15.51
1	Syntactic	続ける	tuzukeru	5,519	539,169	261	1425	6.73	16.21
2		始める	hazimeru	2,983	1,379,861	207	657	6.50	15.55
3		あう	au	2,302	295,787	170	873	6.16	16.87
4		過ぎる	sugiru	3,777	368,408	130	515	5.71	18.74
5		まくる	makuru	86	708,256	32	66	4.56	8.91
6		終わる	owaru	1,884	51,545	31	56	4.50	9.18
7		終える	oeru	503	850,402	24	37	4.31	5.90
8		尽くす	tukusu	687	843,270	26	89	3.72	20.86
9		ぬく	nuku	575	724,584	23	131	3.11	31.33
10		かねる	kaneru	328	1,062,433	18	108	2.82	32.27
11		得る	eru	4,478	362,068	100	1601	0.08	98.77
Means				2,102	653,253	92.91	505.27	4.38	24.96
Standard Deviation				1,786	362,723	83.07	546.90	1.86	24.68

Note. 48 V_2 verbs were selected from the pool of 88 lexical compound and 21 syntactic compound verbs on the baseline of more than 10 token frequencies out of 88,454,573 words in the *Maichini Newspaper.*

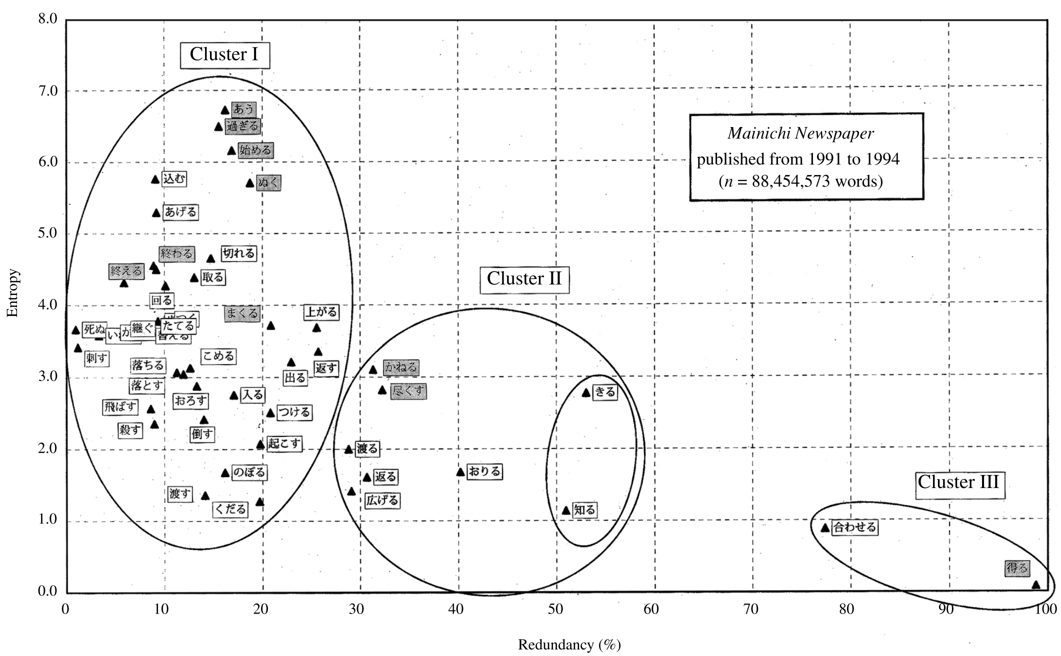


Fig. 2. Plotting and cluster of lexical and syntactic compound V_2 verbs based on entropy and redundancy calculated from the corpus of the *Mainichi Newspaper.*

Note 1: Hierarchical cluster analysis using Ward's method with the square Euclidean distance formed three clusters.

Note 2: 48 lexical and syntactic compound V_2 verbs in this figure were selected from the pool of 109 having more than a token frequency of 10.

Note 3: Verbs in unshadowed boxes are lexical compound verbs ($n = 37$), while verbs in showed boxes are syntactic compound verbs ($n = 11$).

redundancies as depicted in Figure 2. A hierarchical cluster analysis using Ward's method with the square Euclidean distance produced three clusters as indicated by the circles in the figure.

3.2.1. Cluster I

A majority of 38 compound verbs are included in Cluster I. Syntactic compound verbs are likely to be higher in entropy than lexical ones in this cluster. In fact, V_2 verbs of syntactic compounds such as *au* (あう), *sugiru* (過ぎる), *hazimeru* (始める) and *nuku* (ぬく) are very high in entropy, suggesting a wide variety of two verb combinations. Although two V_2 verbs, *komu* (込む) and *ageru* (あげる), are lexical compounds, they also showed a wide range of verb combinations. *Komu* is usually interpreted as "internal movement" as in *nagare* + *komu* (appearing three times, "flow into"), *sosogi* + *komu* (two, "pour into") and *kake* + *komu* (one, "run into"). Although Himeno (1999) classifies these compounds into a type of "internal movement", V_2 *komu* seems to add the meaning "to do thoroughly" to V_1 verbs, such as *oboe* + (19, "teach" +), *kezuri* + (13, "shave" +), *utai* + (five, "sing" +), *yomi* + (three, "read"), and *migaki* + (two, "polish" +) with a very high frequency. Kageyama (1993) pointed out that *komu* adds a certain meaning as does a prefix in English, when it is added to various V_1 verbs to produce a variety of compound verbs, and yet these compounds are lexical. Similarly, the original meaning of *ageru,* "lift" or "elevate", is extended to include "climb up (as in one's status)", "something rising up inside the body", "completion of action" and "emphasis" (Himeno, 1999), resulting in the verb having many combinations with a wide range of V_1 verbs.

3.2.2. Cluster II

Cluster II includes eight V_2 verbs; two syntactic compounds of *kaneru* (かねる) and *tukusu* (尽くす), and six lexical compounds. All these verbs in Cluster II seem to be more redundant than those of Cluster I, especially *kiru* (きる) and *siru* (知る). There are a wide range of V_2 *kiru* combinations with V_1 verbs; 26 compound verbs with *kiru* appear just once and 13 compounds twice. *Kiru* adds three basic meanings of "cut or end", "completion", and "limitation" to V_1 verbs (Himeno, 1999). Although *kiru* is commonly understood to mean "cut", this was seldom used in the newspaper corpus, whereas the meaning "completion" was recurrent, such as *nari* + *kiru* (27, "identify completely"), *mamori* + *kiru* (14, "defend completely"), *watari* + *kiru* (seven, "cross completely"), *uri* + *kiru* (seven, "sell completely"), and *nobori* + *kiru* (seven, “climb completely”). Unlike *kiru*, *siru* only has five different compounds; *ukagai* + *siru* (“ask to know”) appears 19 times while four other compounds appear once or twice. Nevertheless, *kiru* and *siru* share the similar tendency whereby one or few compound verbs appear frequently, while many others appear only once or twice.

3.2.3. Cluster III

The V_2 lexical compound *awaseru* (合わせる) and the V_2 syntactic compound *eru* (得る) were classified in Cluster III. These two V_2 verbs showed a greater pattern of defilement than others. *Awaseru* produced 15 different compounds in the corpus. Among them four lexical compound verbs appeared frequently: 20 times for *kangae* + *awaseru* (“to put thought together”), 11 times for *suri* + *awaseru* (“rub things together”), 10 times for *tunagi* + *awaseru* (“tie things together”), and six times for *terasi* + *awaseru* (“check things with”), while the others appeared once or twice. These four frequently seen compounds could be common expressions in newspapers. V_2 *eru* of syntactic compounds appeared 1601 times, being combined with 100 different V_1 verbs. The compound of *ari* + *eru* (“can do” or “be possible”) showed the highest frequency, counting 865 times. The compound of *nari* + *eru* (“be possible to become”) was the second, appearing 194 times. This diversity might be as the result of the fact that *eru* acts like the auxiliary verb “can” in English.

4. FREQUENCY, ENTROPY AND REDUNDANCY OF COMPOUND VERBS FROM THE NOVEL COLLECTION

The corpus of *Aozora Bunko* collected various novels from the periods of Meiji, Taisho and the beginning of Showa. The collection was gathered from novels and stories written by famous writers, including 18 novels and stories by Kenji Miyazawa (1896–1933), 14 by Nankichi Niimi (1913–1943), seven by Souseki Natsume (1867–1916), 11 by Riichi Yokomitsu (1898–1947), 41 by Kyusaku Yumeno (1889–1936), 105 by Ryunosuke Akutagawa (1892–1927), 19 by Takeo Arishima (1878–1923), 105 by Osamu Dazai (1909–1948), 14 by Doppo Kunikida (1871–1908), 22 by Motojiro Kajii (1901–1932), 12 by Kanoko Okamoto (1889–1939), 28 by Kan Kikuchi (1888–1948) and so on. Thus, the novel corpus, *Aozora Bunko* is considered to be an example of older writings in Japanese. The corpus consisted of a total token frequency, 8,370,720 words. As

Table 2. Frequency, Entropy and Redundancy of Lexical and Syntactic $V_1 + V_2$ Compound Verbs Taken from the *Aozora Bunko* Novel.

#	Type	V_2 verbs		V_2 token frequency	V_1 total token frequency	V_1 type frequency	V_1 & V_2 token frequency	Entropy	Redundancy (%)
		Japanese	Phonetic						
1	Lexical	あげる	ageru	572	25,037	48	92	5.13	8.17
2		かかる	kakaru	603	28,586	46	90	5.08	8.01
3		つく	tuku	658	35,298	41	107	4.78	10.75
4		たてる	tateru	218	7,413	28	57	4.48	6.77
5		出る	deru	1,980	6,355	31	61	4.46	9.95
6		取る	toru	823	4,436	27	45	4.43	6.92
7		きる	kiru	311	109,842	36	69	4.42	14.52
8		のぼる	noboru	460	33,216	37	142	4.39	15.69
9		回る	mawaru	162	24,004	29	79	4.31	11.24
10		歩く	aruku	368	13,012	23	34	4.26	5.91
11		刺す	sasu	150	21,216	20	25	4.21	2.51
12		上がる	agaru	182	45,880	62	174	4.19	6.10
13		返す	kaesu	193	9,719	20	39	4.16	5.28
14		入る	ireru	678	20,161	23	55	4.15	8.28
15		殺す	korosu	358	48,145	19	36	4.04	4.95
16		落ちる	otyiru	210	6,852	15	31	3.55	9.24
17		つける	tukeru	86	20,961	15	30	3.51	10.24
18		合わせる	awaseru	50	30,796	12	15	3.46	3.58
19		倒す	taosu	34	1,603	12	16	3.45	3.69
20		いれる	ireru	287	2,702	12	17	3.29	8.31
21		狂う	kuruu	63	2,581	11	20	3.22	6.87
22		おろす	orosu	67	14,421	10	20	3.18	4.15

(*continued*)

Table 2. (*continued*).

#	Type	V_2 verbs		V_2 token frequency	V_1 total token frequency	V_1 type frequency	V_1 & V_2 token frequency	Entropy	Redundancy (%)
		Japanese	Phonetic						
23		破る	yaburu	74	3,688	10	14	3.18	4.21
24		くだす	kudasu	102	21,056	9	10	3.12	1.51
25		渡る	wataru	157	8,438	12	30	3.11	13.39
26		返る	kaeru	67	7,444	9	42	2.62	17.50
27		込む	komu	281	110,771	81	220	2.13	66.48
28		起こす	okosu	53	5,335	5	15	1.93	16.87
29		消す	kesu	47	706	3	13	1.55	2.30
Means				320	23,092	24.34	55.10	3.72	10.12
Standard Deviation				381	27,019	17.68	50.24	0.89	11.48
1	Syntactic	始める	hazimeru	294	148,945	100	178	6.07	8.67
2		あう	au	268	133,453	85	173	5.72	10.71
3		過ぎる	sugiru	617	67,033	66	132	5.44	10.08
4		続ける	tuzukeru	146	48,805	50	85	5.28	6.51
5		得る	eru	551	159,594	98	297	5.12	22.56
6		かねる	kaneru	84	121,923	44	79	5.02	7.99
7		尽くす	tukusu	20	80,821	10	11	3.28	1.33
8		終わる	owaru	64	17,183	12	24	2.98	16.89
Means				256	97,220	58.13	122.38	4.86	10.59
Standard Deviation				210	47,983	33.24	87.75	1.05	6.08

Note. 37 V_2 verbs were selected from the pool of 88 lexical compound and 21 syntactic compound verbs on the baseline of more than 10 token frequencies in 8,370,720 words from the collection of *Aozora Bunko* novels.

with the newspaper corpus, using a baseline of more than 10 token frequencies, a total of 37 V_2 verbs were selected from the pool of 88 lexical compound and 21 syntactic compound verbs (108 V_2 candidates).

4.1. Comparison of Lexical and Syntactic Compound Verbs

As with the newspaper corpus, the ANOVA on entropy indicated that the lexical compound verbs ($n = 29$, M = 3.72, SD = 0.90) were significantly lower than the syntactic compound verbs ($n = 8$, M = 4.86, SD = 1.12) [$F(1,35) = 9.14$, $p < .01$]. In contrast, the ANOVA on redundancy indicated no significant difference between the lexical compound verbs ($n = 29$, M = 10.12, SD = 11.68) and the syntactic compound verbs ($n = 8$, M = 10.59, SD = 6.50) [$F(1,35) = 0.01$, *n.s.*]. These results also suggested that V_2 verbs of the syntactic compounds were irregularly combined with various V_1 verbs rather than with V_2 verbs of the lexical compounds, while both the lexical and syntactic compound verbs were equally redundant.

4.2. Classification of Lexical and Syntactic Compound Verbs

To investigate individual V_2 tendencies of appearance in the novel corpus further, all the 37 V_2 verbs of lexical and syntactic compounds were plotted based on their entropies and redundancies as in Figure 3. A hierarchical cluster analysis using Ward's method with the square Euclidean distance produced three clusters as indicated by the circles in the figure.

4.2.1. Cluster I

A majority of compound verbs (29 out of 37) were classified into Cluster I. In general, syntactic compound verbs were likely to be higher in entropy than lexical ones in this cluster. The two V_2 verbs of *kakaru* (かかる) and *ageru* (あげる), which combined with V_1 verbs to produce lexical compounds, were also high in entropy. *Ageru* was consistently high in entropy throughout both the corpora of newspapers and novels, while *kakaru* appeared to be high only in the novel corpus. Unlike the 14 different lexical compounds created by V_2 *kakaru* in the newspaper corpus, V_2 *kakaru* combined with 46 different V_1 in the novels. Among them, 28 compounds only appeared once and ten compounds were used twice. However, a variety of compounds with *karaku* seemed to have popular usage in the novels during the Meiji, Taisho, and early Showa periods, these compounds might be falling out of use in the modern Japanese.

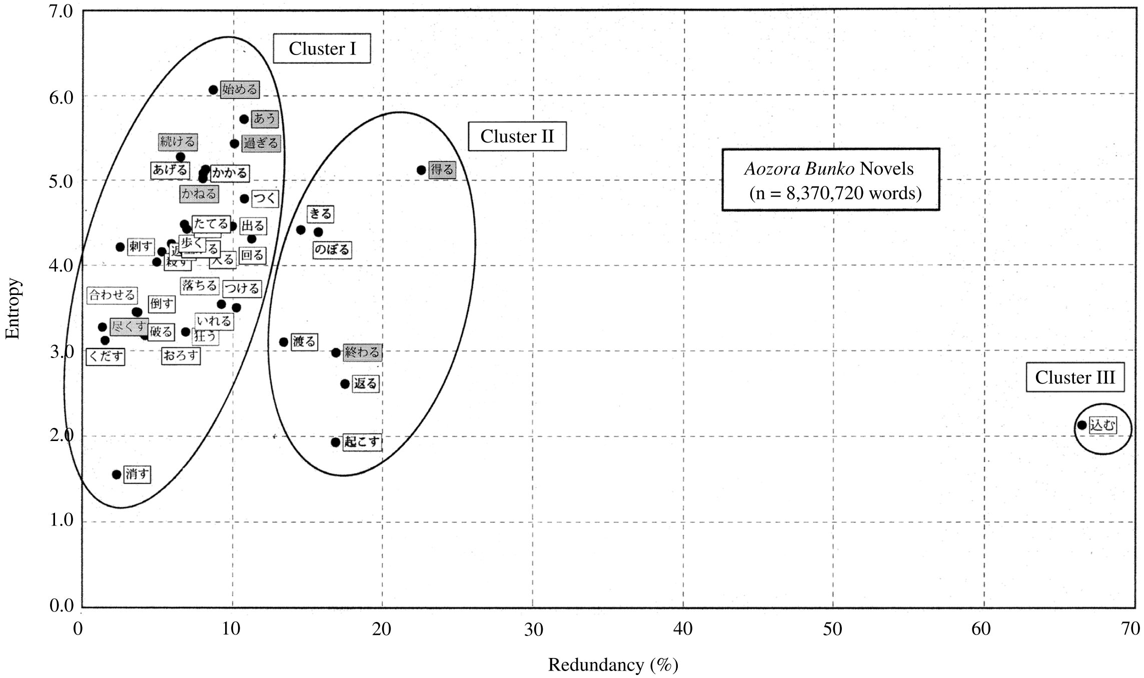


Fig. 3. Plotting and cluster of lexical and syntactic compound V_2 verbs based on entropy and redundancy calculated from the corpus of novels.
Note 1: Hierarchical cluster analysis using Ward's method with the square Euclidean distance formed three clusters.
Note 2: Lexical and syntactic compound V_2 verbs in this figure were selected from the pool of 109 having more than a token frequency of 10.
Note 3: Verbs in shadowed boxes are lexical compound verbs ($n = 29$), while verbs in shadowed boxes are syntactic compound verbs ($n = 8$).

4.2.2. Cluster II

Cluster II includes seven V_2 verbs; two syntactic compounds of *eru* (得る) and *owaru* (終わる), and five lexical compounds. All the verbs in Cluster II were basically more redundant than those of Cluster I. Although V_2 *eru* was extremely high in redundancy in the newspaper corpus as depicted in Figure 2, it appeared only slightly apart from the majority in the novel corpus. Given that *eru* is used like ‘‘can’’ in English, *eru* appeared to be combined with 98 different V_1 verbs, having a total token frequency of 297 times.

4.2.3. Cluster III

Cluster III only consisted of *komu*, being highly redundant. This verb combined with 81 different V_1 verbs, counting up to 220 times of token frequency. The most frequently used compound, *nozoki* + *komu* (‘‘look into’’) appeared 37 times. A wide range of compound expressions with *komu* might be utilized because its accompanying meaning of ‘‘internal movement’’ is suitable for describing various actions and emotions in novels.

5. COMPARING THE LEXICAL AND SYNTACTIC COMPOUND VERBS IN THE CORPORA OF THE NEWSPAPER AND NOVELS

In order to compare compound verbs in the two corpora, 34 V_2 verbs were selected from overlapped items between Figures 1 and 2. The values of entropy and redundancy in the novels were subtracted from those in the newspaper. Based upon these values, they were plotted in Figure 4. Again, a hierarchical cluster analysis using Ward’s method with the square Euclidean distance produced three clusters as indicated by the circles in the figure. The cluster analysis identified that three V_2 verbs deviated far away from the majority of Cluster I. *Komu*, the only V_2 verb in Cluster II, was highly redundant in the novel corpus as seen in Figure 3. V_2 verbs of *awaseru* and *eru* were classified into Cluster III. They were seen to combine with a wide range of V_1 verbs. As explained previously in this study, *awaseru* adds the meaning ‘‘together’’ to V_1 verbs and creates lexical compounds, while *eru* attaches the meaning of ‘‘possibility’’ to V_1 verbs as the auxiliary verb ‘‘can’’ in English does to produce syntactic compounds. Regardless of lexical or syntactic, both V_2 verbs, *awaseru* and *eru* were preferred for use in the newspaper as a function of enlarging

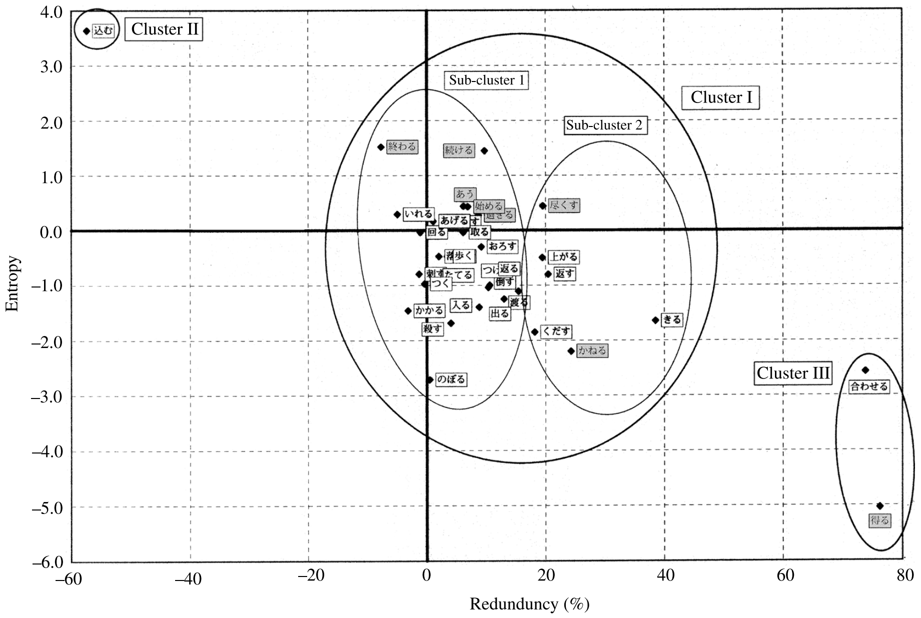


Fig. 4. Plotting and cluster of lexical and syntactic compound V_2 verbs based on differences of entropy and redundancy between the corpus of the newspaper and the novels.

Note 1: Hierarchical cluster analysis using Ward’s method with the square Euclidean distance formed three clusters including two sub-clusters.

Note 2: Compound V_2 verbs ($n = 34$) in this figure were selected from overlapped items between Figures 1 and 2.

Note 3: Verbs in shadowed boxes are lexical compound verbs ($n = 26$), while verbs in unshadowed boxes are syntactic compound verbs ($n = 8$).

Note 4: Differences were calculated from entropy and redundancy of the newspaper corpus subtracted from those of the novel corpus.

simple V_1 expressions. A majority of 31 V_2 verbs were classified into Cluster I with two sub-clusters. These two sub-clusters, however, were located close to each other, so it would be sensible enough to conclude that both lexical and syntactic compound verbs are used with a similar inclination between the newspaper and novel corpora with only a few exceptions.

6. CONCLUSION

The present study investigated lexical and syntactic compound verbs using Shannon's concepts of entropy and redundancy calculated by frequency data from the newspaper and novel corpora. V_2 verbs of lexical compounds were exceeded in number by those of V_2 verbs of syntactic compounds. Using the basis of V_2 token frequency (more than 10), V_2 verbs for syntactic compounds were only counted as 11 types in the newspaper and nine types in the novels, while those for lexical compounds were 37 types in the newspaper and 29 types in the novels. Comparing these V_2 types, syntactic compounds were higher in entropy than lexical ones. This result suggests that V_2 verbs of syntactic compounds are more likely to combine with a wide range of V_1 verbs than those of lexical compounds. As depicted in Figure 1, richness in V_1 and V_2 combinations for syntactic compounds derived from the syntactic structure of a verb phrase, VP_1, constructed with a noun phrase NP-o and a V_1 verb. This enables connection with a variety of V_2 verbs to create a verb phrase VP_2.

Moreover, since *komu* and *ageru*, which both create lexical compounds, add a certain meaning, as does a prefix in English, these two V_2 verbs also showed a wide variety of combinations with V_1. Comparing V_2 verbs in the two corpora, *awaseru* and *eru* indicated a different pattern of entropy and redundancy in the newspaper corpus while *komu* differed in the novel corpus. These exceptional V_2 verbs must reflect characteristics of the corpora. V_2 *eru*, which adds the meaning of "possibility", functions like the auxiliary verb "can" in English, and seems to be a favored expression in the newspapers. In contrast, V_2 *komu*, adds the meaning of "internal movement" like the preposition "into" in English, and appears to be preferred in the novels due to its enrichment of expression of lexical compounds. In general, the present study suggests that both lexical and syntactic compounds were used similarly in the both corpora.

REFERENCES

Himeno, M. (1999). *Fukugoo dooshi no koozoo to imi yoohoo* [Structure of compound verbs and their usage of meanings]. Tokyo: Hitsuji Shobo.

Hori, J. (1979). *What is entropy*? Tokyo: Koodansha Blue Books.

Kaiho, H. (1989). Joohoo-o hakaru – entoropii, joohoo dentatsu-ryoo, joochoodo [Measuring information – Entropy, the amount of transmitted information, and redundancy]. In H. Kaiho (Ed.), *Shinri kyooiku deeta-no kaiseki-hoo 10-koo* [10 lectures on data analysis in psychology and education – application] (pp. 14–26). Tokyo: Fukumura Shuppan.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423 (Part I) and 623–656 (Part II).

Kageyama, T. (1993). *Bunpoo to gokeisei* [Grammar and word formation]. Tokyo: Hitsuji Shoboo.

Kageyama, T. (1999a). Keitairon to imi [Morphology and meaning]. Tokyo: Kuroshio Shuppan.

Kageyama, T. (1999b). Word formation. In N. Tsujimura (Ed.), *The handbook of Japanese linguistics* (pp. 297–325). Malden, Massachusetts: Blackwell Publishers.