2019 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
SCIENCE, TECHNOLOGY & ENGINEERING, ARTS, MATHEMATICS & EDUCATION JUNE 5 - 7, 2019
HAWAII PRINCE HOTEL WAIKIKI, HONOLULU, HAWAII

INVESTIGATION OF WORDS IN JAPANESE CLOSED CAPTION TV CORPUS

MOCHIZUKI, HAJIME
INSTITUTE OF GLOBAL STUDIES
TOKYO UNIVERSITY OF FOREIGN STUDIES
FUCHU, TOKYO
JAPAN

Investigation of Words in a Japanese Closed Caption TV Corpus

Hajime Mochizuki
Institute of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan

Abstract: For Japanese learners, we describe the specific details of TV program vocabulary and investigate what kinds of words are necessary for understanding the contents of TV scripts. We use our closed caption TV corpus over 1 billion words in size for the investigation of vocabulary. In this paper we will show different word statistics from various viewpoints such as the difference in years and the difference in parts-of-speech.

Introduction

Corpus-based academic word lists have received much attention in last few decades (Gardner and Davies, 2013). The academic words refer to words that are used for building knowledge and conceptual understanding in academic texts or dialogues, including specific domains such as politics, science, and technology. Academic words can also be said to be a limited vocabulary for reading and writing academic contents.

However we can also consider a general vocabulary for language learners to encompass a wide variety of contents. In this study, we will focus on the investigation of words in TV programs for the understanding of TV contents. A wide variety of topics are included in Japanese TV programs, including documentaries, daily information reports, news, sports, dramas, and so forth. We therefore think that watching TV programs is one of the most effective methods of learning Japanese and understanding Japanese culture.

When we investigate words in TV programs, we need a large-scale TV script corpus. There exists little research on words in TV scripts due to a lack of language resources. On the other hand, we have been building a large-scale spoken language corpus from Japanese closed caption TV (CCTV) data transmitted through digital terrestrial broadcasting since December 2012 (Mochizuki and Shibano, 2014). The size of our corpus has reached over 340,000 TV programs, and over 119 million sentences as of December 2018. Corpora have become the most important resources for research and applications related to natural language, and a variety of research and applications for corpus-based computational linguistics, knowledge engineering, and language education have been reported in recent years (Flowerdew, 2011, Newman, Baayen, and Rice, 2011).

In this paper, we describe the specific details of TV program vocabulary and investigate what kinds of words are necessary for understanding the contents of TV scripts.

Part-Of-Speech Statistics of Closed Caption TV Corpus

Here, we provide some statistical facts regarding our corpus. We have been building a large-scale spoken language corpus from Japanese closed caption TV (CCTV) data. The size

of our corpus has reached over 340,000 TV programs, over 119 million sentences, and about 1.3 billion words. Table 1 shows the number of TV programs, sentences, and words recorded every year from 2013 to 2018. This research is based on our corpus which includes a total of 340,418 programs, 119,431,847 sentences, or 1,291,171,749 words as shown in Table 1.

Table 1: The numbers of TV programs, sentences, and words on CCTV corpus

Year	Programs	Sentences	Words
2013	44,110	15,482,501	162,690,805
2014	54,091	19,341,947	206,792,273
2015	60,330	20,537,381	223,194,770
2016	58,850	20,756,005	228,091,667
2017	61,873	21,744,151	235,951,570
2018	61,164	21,569,862	234,450,664
Total	340,418	119,431,847	1,291,171,749

Table 2 shows the two types of frequencies and ratios by part-of-speech categories. The first frequency and ratio are the number of tokens listed in third column and its ratio is listed in the fourth column, and the second frequency and ratio are the number of types listed in the fifth column and its ratio is listed in the sixth column.

Table 2: POS of the closed caption TV corpus for six-years.

Part-of-speech (POS)	Token	Ratio(%)	Type	Ratio(%)
nouns (Noun)	435,353,188	33.72	809,316	94.09
particle (Part)	328,840,149	25.47	207	0.02
verb (Verb)	160,732,645	12.45	11,799	1.37
symbol (Sym)	133,321,611	10.33	87	0.01
auxiliary verb (Aux)	126,902,024	9.83	36	0.00
adverb (Adv)	33,081,802	2.56	2,679	0.31
adjective (Adj)	20,950,179	1.62	1,423	0.02
prenominal adjective (PreAdj)	14,200,020	1.10	121	0.01
Interjection (Interj)	13,898,350	1.08	1,531	0.17
conjunction (Conj)	11,708,583	0.91	160	0.01
prefix (Prefix)	8,642,924	0.67	178	0.02
filler (Fil)	2,977,970	0.23	18	0.00
Other	562,304	0.04	32,610	3.79
Total	1,291,171,749		860,165	

From the viewpoint of the tokens of words, the sum of the frequencies for the top five most frequent part-of-speech categories is over 91% of all tokens. The ratio for the sum of the word frequencies for nouns is about 33.7%, followed in order by particles, verbs, symbols, and auxiliary verbs which are 25.5%, 12.5%, 10.3%, and 9.8%, respectively in Table 2. These proportions of part-of-speech categories suggest a real distribution of these categories in a Japanese sentence in the CCTV corpus. On the other hand, from the viewpoint of vocabularies, a large amount is occupied by nouns and verbs. The ratios for the types of nouns and verbs are

over 94% and 1.4%, respectively, though the ratios of particles, symbols, and auxiliary verbs are only 0.02%, 0.01%, and less than 0.01%, respectively. This indicates that particles, symbols, and auxiliary verbs should be given priority in learning Japanese because they can cover a wide range of the Japanese CCTV corpus by using smaller vocabularies.

Word Statistics of the Closed Caption TV Corpus

The top 50 most frequent word statistics of our CCTV corpus for six years from 2013 to 2018 are listed in Table 3. In Table 3, Rank indicates the order based on the total word frequency over the course of six years. POS indicates the part-of-speech for each Word. Frequency indicates the number of occurrences for each word in the corpus. Column "Cov." indicates the coverage. To calculate the coverage, the sum of the word frequencies for each rank is divided by the total number of word frequencies. The sixth to eleventh columns indicate the orders based on one-year frequencies from 2013 to 2018. Table 3 shows that the distribution of the top 50 most frequent words is almost the same during the last six years. This suggests that there is a certain order of frequently used words in Japanese TV.

Table 3: Top 50 words based on their frequency on the closed caption TV corpus for six-years.

Rank	Word	POS	Frequency	Cov.	2013	2014	2015	2016	2017	2018
1	。(-)	Sym.	81,234,123	6.3	1	1	1	1	1	1
2	の (*no*)	Part	44,375,961	9.7	2	2	2	2	2	2
3	て (*te*)	Part	34,489,210	12.4	3	3	3	3	3	3
4	た (*ta*)	Aux	33,159,406	15.0	4	4	4	4	4	4
5	に (*ni*)	Part	32,650,684	17.5	5	5	5	5	5	5
6	が (*ga*)	Part	30,469,624	19.9	7	6	6	6	6	6
7	は (*wa*)	Part	30,004,327	22.2	6	7	7	7	7	7
8	を (*wo*)	Part	26,020,347	24.2	8	8	8	8	8	8
9	です (*desu*)	Aux	25,112,325	26.1	9	9	10	10	10	9
10	、(-)	Sym	24,441,310	28.0	11	10	9	9	9	10
11	ます (*masu*)	Aux	23,646,746	29.9	10	11	11	11	11	11
12	だ (*da*)	Aux	21,850,673	31.6	12	12	12	12	12	12
13	する (*suru*)	Verb	20,846,268	33.2	13	13	13	13	13	13
14	と (*to*)	Part	17,490,763	34.5	14	14	14	14	14	14
15	で (*de*)	Part	16,891,975	35.8	15	15	15	15	15	15
16	ね (*ne*)	Part	12,898,557	36.8	18	16	16	16	16	16
17	… (-)	Sym	12,102,287	37.8	19	19	17	17	17	17
18	? (-)	Sym	11,751,625	38.7	16	17	18	21	20	22
19	! (-)	Sym	11,751,058	39.6	17	18	19	19	18	19
20	ん (*n*)	Noun	11,458,729	40.5	20	20	20	20	19	21
21	いる (*iru*)	Verb	11,340,285	41.4	23	22	21	18	21	20
22	も (*mo*)	Part	11,255,592	42.2	22	21	22	22	22	18
23	か (*ka*)	Part	11,109,784	43.1	21	23	23	23	23	23
24	ない (*nai*)	Aux	9,806,878	43.8	24	24	24	24	24	24
25	よ (*yo*)	Part	7,955,395	44.5	25	25	25	25	25	25

26	から (*kara*)	Part	7,520,897	45.0	26	26	26	26	26	26
27	の (*no*)	Noun	7,256,172	45.6	27	27	27	27	27	27
28	さん (*san*)	Noun	6,423,951	46.1	28	28	28	29	28	29
29	ある (*aru*)	Verb	6,288,105	46.6	29	29	29	28	29	28
30	なる (*naru*)	Verb	5,964,562	47.1	30	30	30	30	30	30
31	れる (*reru*)	Verb	5,677,213	47.5	31	31	31	31	31	31
32	こと (*koto*)	Noun	5,321,498	47.9	33	32	32	32	32	32
33	〜 (-)	Sym	5,069,433	48.3	34	34	33	34	33	33
34	という (*toiu*)	Part	4,901,157	48.7	39	36	34	33	34	34
35	てる (*teru*)	Verb	4,746,615	49.0	32	33	36	36	36	36
36	この (*kono*)	PreAdj	4,737,103	49.4	35	35	35	35	35	35
37	これ (*kore*)	Noun	4,536,774	49.8	36	37	37	37	37	37
38	人 (*nin*)	Noun	4,132,234	50.1	38	38	38	38	39	38
39	う (*u*)	Aux	4,113,495	50.4	37	39	39	39	38	39
40	お (*o*)	Preffix	3,650,778	50.7	41	40	40	40	40	40
41	「	Sym	3,391,251	50.9	40	41	41	42	42	42
42	って(*tte*)	Part	3,308,965	51.2	43	42	42	41	41	41
43	思う(*omou*)	Verb	3,098,686	51.4	47	44	43	43	43	43
44	な (*na*)	Part	3,050,961	51.7	44	45	44	44	44	44
45	」	Sym	3,043,502	51.9	42	43	45	48	46	47
46	その (*sono*)	PreAdj	2,935,671	52.1	48	47	46	45	45	45
47	何 (*nani*)	Noun	2,891,442	52.4	45	46	47	50	48	50
48	いい (*ii*)	Adj	2,869,048	52.6	46	48	48	49	47	49
49	くる (*kuru*)	Verb	2,848,637	52.8	50	49	49	47	49	46
50	1	Noun	2,841,394	53.0	49	50	50	46	50	48

In the first place is the "。" symbol, which indicates the period of a sentence in Japanese. The words *no*, *te, ni*, *wa*, *ga*, *wo*, *to*, *de*, *ne*, *mo*, *ka*, *yo*, *kara*, *toiu*, *tte*, and *na* in the list are particles, which are very important discourse makers for understanding the meaning of a Japanese sentence (Tsujimura, 1996). The words *ta*, *desu*, *masu*, *da*, *nai*, and *u* in the list are auxiliary verbs. In Japanese, auxiliary verbs can refer to either inflecting suffixes or auxiliary verbs (verb suffixes that are verbs on their own).[1]

The list also includes verbs and nouns. However, all of verbs except for *omou*, namely, *suru*, *iru*, *aru*, *naru*, *reru*, *teru*, and *kuru*, are formal verbs that do not represent substantive meaning, but are close to auxiliary verbs. Only one noun in all eight nouns is an Arabic numeral and the remaining seven nouns do not represent substantive meaning, but are formal nouns. Among them, *kore* is akin to a demonstrative pronoun, *n* and *no* are akin to particles, *san* and *nin* are akin to postfixes, and *koto* refers to a matter or a thing. As shown in Table 3, there are no nouns and verbs that play a role as content words in the top 50 most frequently used words. The following 50 words ranked from 51st to 100th are listed in Table 4.

[1] Wiktionary (https://en.wiktionary.org/wiki/Appendix:Japanese_auxiliary_verbs)

Table 4: Words list based on their frequency for six-years (ranked from 51st to 100th).

Rank	Word	POS	Frequency	Cov.	2013	2014	2015	2016	2017	2018
51	そう (*sou*)	Adv	2,625,344	53.2	51	51	51	51	51	51
52	よう (*you*)	Noun	2,493,432	53.4	53	53	52	52	52	53
53	それ (*sore*)	Noun	2,458,647	53.6	52	52	53	53	53	55
54	2	Noun	2,433,824	53.8	54	54	54	54	54	52
55	けど (*kedo*)	Part	2,367,703	54.0	55	55	55	55	55	57
56	やる (*yaru*)	Verb	2,254,408	54.2	58	57	57	57	57	59
57	いう (*iu*)	Verb	2,230,783	54.3	64	60	56	56	56	58
58	ない (*nai*)	Adj	2,215,108	54.5	59	58	58	58	58	61
59	じゃ (*jya*)	Part	2,207,206	54.7	57	56	59	60	59	62
60	言う (*iu*)	Verb	2,168,411	54.8	56	59	60	61	60	63
61	いく (*iku*)	Verb	2,116,072	55.0	65	61	61	59	61	60
62	見る (*miru*)	Verb	2,041,819	55.2	63	63	63	62	63	66
63	ちょっと (*chotto*)	Adv	2,025,767	55.3	66	65	62	63	62	64
64	ん (*n*)	Aux	1,964,997	55.5	62	62	65	65	68	68
65	事 (*koto*)	Noun	1,947,283	55.6	61	64	64	64	64	78
66	たい (*tai*)	Aux	1,919,483	55.8	67	67	67	66	67	69
67	もう (*mou*)	Adv	1,902,910	55.9	69	69	66	67	65	67
68	はい (*hai*)	Interj	1,896,690	56.1	68	68	68	70	66	65
69	私 (*watashi*)	Noun	1,844,528	56.2	60	66	72	71	71	76
70	まで (*made*)	Part	1,831,891	56.4	70	70	69	69	70	71
71	ここ (*koko*)	Noun	1,830,009	56.5	71	71	71	68	69	70
72	や (*ya*)	Part	1,782,156	56.6	72	72	70	73	72	72
73	中 (*tyuu*)	Noun	1,748,528	56.8	74	73	74	74	73	74
74	できる (*dekiru*)	Verb	1,718,792	56.9	77	74	76	76	76	73
75	今 (*ima*)	Noun	1,710,638	57.0	75	75	75	72	75	79
76	年 (*nen*)	Noun	1,705,080	57.2	73	77	73	77	77	77
77	3	Noun	1,698,552	57.3	78	76	77	75	74	75
78	どう (*dou*)	Adv	1,614,758	57.4	76	78	78	78	79	83
79	すごい (*sugoi*)	Adj	1,580,937	57.5	81	83	81	79	78	80
80	られる (*rareru*)	Verb	1,545,234	57.7	80	80	79	80	81	85
81	ば (*ba*)	Part	1,541,813	57.8	79	79	82	82	80	86
82	そして (*soshite*)	Conj	1,519,768	57.9	86	81	83	81	83	82
83	もの (*mono*)	Noun	1,519,590	58.0	83	82	80	83	82	87
84	日 (*nichi*)	Noun	1,457,116	58.1	84	84	85	86	84	89
85	ので (*node*)	Part	1,443,310	58.2	96	88	89	84	85	81
86	=	Sym	1,436,199	58.4	82	106	103	94	86	54
87	など (*nado*)	Part	1,430,508	58.5	90	87	84	85	88	91
88	出る (*deru*)	Verb	1,424,408	58.6	85	85	87	88	90	90
89	とか (*toka*)	Part	1,418,269	58.7	89	90	86	87	87	88
90	方 (*hou*)	Noun	1,377,348	58.8	88	86	88	90	92	93
91	こちら (*kochira*)	Noun	1,356,513	58.9	94	92	90	89	91	95
92	たち (*tachi*)	Noun	1,339,035	59.0	87	89	91	91	96	96
93	時 (*toki*)	Noun	1,308,932	59.1	92	91	92	93	94	97

94	だけ (*dake*)	Part	1,288,170	59.2	93	93	93	95	95	98
95	そう (*sou*)	Noun	1,257,600	59.3	97	96	98	97	93	94
96	前 (*zen*)	Noun	1,246,547	59.4	95	95	95	96	99	99
97	者 (*sha*)	Noun	1,242,321	59.5	100	97	94	92	98	101
98	行く (*iku*)	Verb	1,216,778	59.6	91	94	97	99	100	104
99	みる (*miru*)	Verb	1,193,194	59.7	105	99	96	98	97	103
100	日本 (*nippon*)	Noun	1,144,028	59.8	109	98	99	100	106	100

There are ten categories in the POS column of Table 4, which include nine particles (labeled as "Part"), one symbol ("Sym"), two auxiliary verbs ("Aux"), twenty nouns ("Noun"), ten verbs ("Verb"), two adjectives ("Adj"), four adverbs ("Adv"), one interjection ("Interj"), and one conjunction ("Conj"). As in Table 3, many words in Table 4 are function words such as particles and auxiliary verbs. Although 20 nouns are included, many of them are not content words but are formal nouns used to express references or anaphora, units, and time and date. Only two nouns, *watashi* meaning I and *nippon* meaning Japan, are content words. Similarly, ten verbs are included though seven of them are formal verbs, and only three words, *deru* meaning go out, *miru* meaning see, and *iku* meaning go, are normal verbs. The appearance tendency of words in the second group (Table 4) is almost the same as in the first group (Table 3). The sum of the word frequencies until the 100th word is 771,775,913, although our corpus has 860,165 types and 1,291,207,045 tokens. Therefore, we can say that a small number of words, about 0.01% of all vocabularies, covers more than 59.8% of the 1.29 billion tokens in the entire corpus.

Table 5 shows the coverages according to the words list based on their frequencies. The cover ratios for the total frequencies of words against the top 100, 500, 1,000, 2,500, 5,000, and 10,000 types of words are 59.8%, 73.4%, 78.9%, 86.1%, 90.9%, and 94.8%, respectively. Furthermore, the coverages of the top 15,000, 20,000 and 25,000 are 96.5%, 97.5%, and 98.1%, respectively. Roughly speaking, if Japanese language learners memorize the top 10,000 words on the list, they can obtain the vocabularies necessary for understanding over 94.8% of all contents used by Japanese TV programs. Moreover, if they memorize the top 25,000 words, they can obtain over 98.1% vocabularies. The ratios for the total 860,165 types of words against the top 10,000 and 25,000 types of words are only 1.16% and 2.9%, respectively. This suggests that an accurate word list that is based on a true distribution can be acquired based on real frequencies on the large-scale corpus. The acquired list will cover a wide range of TV vocabularies by a limited size of vocabularies, and will become a profitable learning material for Japanese learners.

Table 5: Frequency Based Ranking and Coverages

Rank	Word	Frequency	Sum of frequencies	Coverage
1	。(-)	81,234,123	81,234,123	6.3%
100	日本 (*nippon*)	1,144,028	771,775,913	59.8%
500	15	205,724	947,863,481	73.4%
1,000	やってくる (*yattekuru*)	104,546	1,019,316,032	78.9%
2,500	バック (*back*)	38,711	1,112,021,103	86.1%
5,000	南北 (*nanboku*)	16,287	1,174,107,480	90.9%

10,000	菜の花 (*nanohana*)	6,072	1,223,738,061	94.8%
15,000	かたまる (*katamaru*)	3,201	1,245,828,159	96.5%
20,000	大ぶり (*ooburi*)	1,966	1,258,402,180	97.5%
25,000	ドネツク (donetsuku)	1,305	1,266,427,666	98.1%
860,165	ぎがひどく (gigahidoku)	1	1,291,171,749	100.0%

From the investigation in this study, we are able to set the basic goal to memorize the top 10,000 most frequent words for learners who want to understand Japanese TV programs. The top 10,000 most frequent words that learners will memorize consist of 7,517 nouns, 1,386 verbs, 371 adverbs, 210 adjectives, 128 particles, 108 interjections, 92 prefixes, 75 conjunctions, 39 prenominal adjectives, 31 auxiliary verbs, 23 symbols, 13 fillers, and 7 other symbols.

Word Statistics for each Part-Of-Speech

Frequency Ranking of Nouns

Table 6 shows the top 50 nouns ordered by frequency in the CCTV corpus. In Table 6, R1 and R2 indicate the rankings of nouns and all words, respectively. There are many words for expressing oneself or other people, namely, *watashi* (means I), *jiubun* (oneself), *ore* (I), *boku* (I), *senshu* (sport player), and *minna* (everyone). These nouns are marked by *1 in Table 6.

There are also many formal nouns that do not represent substantive meanings as in Table 3. They consist of nouns that play roles akin to demonstrative pronouns (*kore*, *sore*, *koko*, and *kochira*) marked by *2, particles (*n*, *no*, *you*, *sou*, *tame*, *sa*, and *mitai*) marked by *3, postfixes (*san*, *nin*, *chuu*, *kon*, *nen*, *nichi*, *tachi*, *ji*, *zen*, *sha*, *teki*, *me*, *jikan*, and *chan*) marked by *4, and expressing matters or things (*koto*, *nani*, *mono*, and *tokoro*) marked by *5. As with other nouns, the top 50 most frequently used nouns list has a few normal nouns (*kyo* and *onegai*) marked by *6, one proper noun (*Nippon* meaning Japan) marked by *7, five Arabic numerals, and some symbols.

Table 6: Frequent nouns (top 50)

R1	R2	Word	Frequency	R1	R2	Word	Frequency
1	20	ん (*n*)*3	11,458,729	26	96	前 (*zen*, former)*4	1,246,547
2	27	の (*no*)*3	7,256,172	27	97	者 (*sha*, person)*4	1,242,321
3	28	さん (*san*, Mr.Mis.)*4	6,423,951	28	100	日本 (*nippon*, Japan)*7	1,144,028
4	32	こと (*koto*, thing)*5	5,321,498	29	104	ため (*tame*)*3	1,073,490
5	37	これ (*kore*, this)*2	4,536,774	30	108	的 (*teki*)*4	1,051,144
6	38	人 (*nin*, person)*4	4,132,234	31	114	ところ (*tokoro*)*5	1,011,902
7	47	何 (*nani*, what)*5	2,891,442	32	115	今日 (*kyou*, today)*6	1,006,963
8	50	1	2,841,394	33	117	さ (*sa*)*3	992,561
9	52	よう (*yo*)*3	2,493,432	34	118	4	972,250
10	53	それ (*sore*, it)*2	2,458,647	35	120	自分 (*jibun*, oneself) *1	953,654
11	54	2	2,433,824	36	122	≫	935,722

12	65	事 (*koto*, thing)*5	1,947,283	37	124	一	921,313
13	69	私 (*watashi*, I)*1	1,844,528	38	126	あと (*ato*, after)*4	886,056
14	71	ここ (*koko*, here)*2	1,830,009	39	128	目 (*me*, ordinal number)*4	877,777
15	73	中 (*chuu*)*4	1,748,528	40	129	5	871,306
16	75	今 (*kon*)*4	1,710,638	41	131	とき (*toki*)*5	833,089
17	76	年 (*nen*, year)*4	1,705,080	42	134	時間 (*jikan*, time)*4	818,491
18	77	3	1,698,552	43	136	俺 (*ore*, I)*1	806,461
19	83	もの (*mono*, thing)*5	1,519,590	44	145	選手 (*senshu*, player) *1	777,886
20	84	日 (*nichi*, day)*4	1,457,116	45	148	僕 (*boku*, I) *1	769,139
21	90	方 (*kata*)*4	1,377,348	46	149	ちゃん (*chan*)*4	750,302
22	91	こちら(*kochira*, here)*2	1,356,513	47	151	みたい (*mitai*, look like)*3	746,782
23	92	たち (*tachi*)*4	1,339,035	48	154	みんな(*minna*, everyone) *1	732,546
24	93	時 (*ji*, hour)*4	1,308,932	49	160	お願い (*onegai*, wish)*6	708,482
25	95	そう (*sou*, so)*3	1,257,600	50	161	!?	705,727

Table 7 shows the following 50 nouns from 51st to 100th ordered according to their frequency in the CCTV corpus. As in Table 6, there are many words for expressing oneself or other people, namely, *kun* or *kimi* (you), *minasan* (everyone), *sensei* (teacher), *omae* (you), *anata* (you), and *dare* (who). These nouns are marked by *1 in Table 7. Furthermore, the list includes many formal nouns that do not represent substantive meanings. They consist of nouns that play roles akin to demonstrative pronouns (*soko*, *are*, and *doko*) marked by *2, particles (*wake*, *mon*, *nan*, and *hou*) marked by *3, and postfixes (*ken*, *kai*, *sai*, *gatsu*, *shi*, *fun*, *en*, *ke*, *man*, *do*, *jyou*, *wa*, *ijou*, *saki*, *go*, *sei*, and *ra*) marked by *4. As with other nouns, the list has a relatively small number of normal nouns (*kanji*, *konkai*, *kyou*, *ki*, *mise*, *sekai*, *josei*, *mondai*, *issho*, *saigo*, *daijobu*, and *otoko*) marked by *6, one proper noun (*Tokyo*) marked by *7, three Arabic numerals, and some symbols.

Table 7: Frequent nouns (from 51 to 100)

R1	R2	Word	Frequency	R1	R2	Word	Frequency
51	163	県 (*ken*, prefecture)*4	699,306	76	212	度 (*do*, degree)*4	535,417
52	165	感じ (*kanji*, feeling)*6	689,577	77	213	上 (*jyou*)*4	533,876
53	166	回 (*kai*, times)*4	682,981	78	215	世界 (*sekai*, world)*6	528,936
54	167	わけ (*wake*, reason)*3	677,073	79	218	話(*wa*, episode no.)*4	519,436
55	169	歳 (*sai*, years old)*4	668,669	80	223	どこ (*doko*, where)*2	513,489
56	171	月 (*gatu*, month)*4	663,593	81	226	手 (*te*, hand)*6	492,823
57	173	そこ (*soko*, there)*2	662,794	82	227	お前 (*omae*, you)*1	492,374
58	176	今回 (*konkai*, this time)*6	651,086	83	230	女性(*josei*, female)*6	479,824
59	179	君 (*kun*)*1	630,668	84	231	問題(*mondai*, problem)*6	477,596
60	180	市 (*shi*, city)*4	629,383	85	232	以上 (*ijou*, over)*4	477,292
61	182	10	624,741	86	233	あなた (*anata*, you)*1	475,672
62	183	分 (*fun*, minute)*4	622,373	87	235	先 (*saki*, beyond)*4	470,231
63	187	きょう (*kyou*, today)*6	614,248	88	236	誰 (*dare*, who)*1	469,967
64	188	円 (*en*, yen)*4	612,610	89	238	後 (*go*, after)*4	467,250
65	194	気 (*ki*, feeling)*6	591,187	90	241	もん (*mon*)*3	460,638

66	196	東京 (*Tokyo*)*7	585,621	91	242	性 (*sei*, nature)*4	459,047
67	197	家 (*ke*, family)*4	583,079	92	243	7	450,802
68	198	皆さん (*minasan*, everyone)*1	572,839	93	245	なん (*nan*)*3	446,283
69	199	一 (*ichi*, one)	568,966	94	247	一緒 (*issho*, together)*6	441,214
70	200	万 (*man*, ten thousand)*4	568,885	95	249	ほう (*hou*)*3	440,896
71	204	）	548,460	96	251	最後 (*saigo*, final)*6	438,943
72	206	6	544,219	97	252	うち (*uchi*, inside)*3	438,567
73	209	あれ (*are*, that)*2	538,867	98	253	大丈夫 (*daijoubu*, all right)*6	435,148
74	210	先生 (*sensei*, teacher)*1	538,799	99	254	男 (*otoko*, male)*6	434,428
75	211	店 (*mise*, shop)*6	538,184	100	255	ら (*ra*)*4	434,397

In the top 100 most frequently used nouns, the most frequent noun type is formal nouns, followed in order by normal nouns, nouns to express oneself or other people, numerals, other symbols, and proper nouns. These types of nouns are 58, 15, 12, 9, 4, and 2, respectively. The results show the following features of nouns in the Japanese CCTV corpus: (1) large portions of the uppermost frequent nouns are formal nouns which are used as functional words; (2) many variations of nouns are used frequently to express oneself or other people; (3) proper nouns hardly appear in the top 100 list of nouns.

The major proper nouns that follow *Nippon* and *Tokyo* are *Amerika* (meaning America, ranked 106th in nouns, 272nd in all words), *chugoku* (China, 179, 423), *Orimpikku* (Olympic games, 218, 485), *Kankoku* (Korea, 350, 712), *Osaka* (377, 754), *Torampu* (Trump, 435, 846), *Rosia* (Russia, 507, 965), *Okinawa* (517, 984), *Tanaka* (663, 1192), *Furansu* (French, 670, 1205), and *Suzuki* (676, 1205).

The major normal nouns after the 100th are *basho* (location, 107, 273), *sigoto* (work, 108, 278), *suki* (like, 109, 279), *jiken* (accident, case, 110, 280), *ame* (rain, 111, 281), *ichiban* (first place, 114, 290), *kankei* (relation, 116, 295), *mizu* (water, 117, 300), *kuruma* (car, 119, 303), *keisatu* (police, 121, 308), *joho* (information, 123, 315), *yougi* (suspicious, 124, 316), *hituyo* (necessary, 125, 317), *kodomo* (children, 127, 319), *koe* (voice, 133, 331), *ryouri* (cooking, 134, 334), *kuni* (country, 138, 345), *tiimu* (team, 139, 347), *dansei* (male, 141, 352), and *jidai* (age, 144, 358).

Frequency Ranking of Verbs

Table 8 shows the top 50 verbs ordered by their frequency in the CCTV corpus. R1 and R2 indicate the rankings of verbs and all words, respectively. There are many formal verbs that do not represent substantive meanings as with Table 3. They consist of 23 basic verbs, namely, *omou*, *iu*, *miru*, *deru*, *iku*, *hairu*, *kuru*, *wakaru*, *taberu*, *tuzuku*, *tukau*, *kiku*, *tukuru*, *okonau*, *matu*, *siru*, *chigau*, *ireru*, *kangaeru*, *dasu*, *mieru*, *matu*, and *toru* marked by *1, and 27 verbs spelled by hiragana letters that play roles akin to auxiliary verbs, namely, *suru*, *iru*, *aru*, *naru*, *reru*, *teru*, *kuru*, *yaru*, *iu*, *iku*, *dekir*u, *rareru*, *miru*, *seru*, *kureru*, *chau*, *kudasaru*, *simau*, *ku*, *itadaku*, *wakaru*, *morau*, *oru*, *ikeru*, *kakeru*, *tukeru*, and *sireru* marked by *2.

Table 8: Top 50 Frequent Verbs

R1	R1	Word	Frequency	R1	R2	Word	Frequency
1	13	する (*suru*, do)*2	20,846,268	26	139	食べる (*taberu*, eat)*1	803,477
2	21	いる (*iru*, be, exist)*2	11,340,285	27	140	しまう (*shimau*)*2	801,756
3	29	ある (*aru*, be, exist)*2	6,288,105	28	141	続く(*tsuzuku*, continue)*1	800,993
4	30	なる (*naru*, become)*2	5,964,562	29	142	使う (*tsukau*, use)*1	800,150
5	31	れる (*reru*)*2	5,677,213	30	150	聞く (*kuku*, listen)*1	749,981
6	35	てる (*teru*)*2	4,746,615	31	152	作る (*tsukuru*, make)*1	740,297
7	43	思う (*omou*, think)*1	3,098,686	32	155	行う (*okonau*, do)*1	732,365
8	49	くる (*kuru*, come)*2	2,848,637	33	157	く (*ku*)*2	724,196
9	56	やる (*yaru*, do, send)*2	2,254,408	34	158	持つ (*motsu*, bring)*1	723,010
10	57	いう (*iu*, say)*2	2,230,783	35	168	いただく(*itadaku*, take)*2	675,891
11	60	言う (*iu*, say)*1	2,168,411	36	170	わかる (*wakaru*, understand)*2	664,314
12	61	いく (*iku*, go)*2	2,116,072	37	172	知る (*shiru*, know)*1	663,107
13	62	見る (*miru*, look)*1	2,041,819	38	181	違う(*chigau*, different)*1	625,152
14	74	できる (*dekiru*, can)*2	1,718,792	39	184	入れる (*ireru*, put in)*1	621,632
15	80	られる (*rareru*)*2	1,545,234	40	189	もらう (*morau*, take)*2	622,516
16	88	出る(*deru*, get out)*1	1,424,408	41	195	考える(*kangaeru*, think)*1	587,335
17	98	行く (*iku*, go)*1	1,216,778	42	225	出す (*dasu*, put out)*1	493,910
18	99	みる (*miru*, look)*2	1,193,194	43	239	おる (*oru*)*2	463,330
19	106	入る (*hairu*, enter)*1	1,052,119	44	240	見える (*mieru*, see)*1	462,077
20	107	来る (*kuru*, come)*1	1,051,827	45	250	待つ (*matsu*, wait)*1	439,241
21	109	せる (*seru*)*2	1,035,964	46	263	いける (*ikeru*)*2	425,248
22	110	くれる (*kureru*, give)*2	1,033,884	47	267	かける (*kakeru*)*2	414,161
23	111	ちゃう (*chau*, defferent, do)*2	1,030,249	48	269	つける (*tsukeru*)*2	403,912
24	119	くださる (*kudasaru*, give)*2	972,113	49	284	しれる (*shireru*)*2	382,306
25	123	分かる (*wakaru*, understand)*1	934,619	50	285	取る (*toru*, put)*1	379,087

Table 9 shows the following 50 verbs from 51st to 100th ordered by their frequency in the CCTV corpus. The list consists of 40 normal verbs marked by *1, and 10 verbs spelled by hiragana letters that play roles akin to auxiliary verbs, namely, *yoru*, *tuku*, *toru*, *kudasaru*, *oku*, *kakaru*, *ageru*, *deru*, *nasaru*, and *sugiru* marked by *2.

Table 9: Frequent verbs (from 51st to 100th)

R1	R2	Word	Frequency	R1	R2	Word	Frequency
51	286	受ける(*ukeru*, receive)*1	347,047	76	442	あげる (*ageru*, give)*2	231,311
52	287	続ける (*tsuzukeru*, continue)*1	376,473	77	445	教える (*oshieru*, teach)*1	228,648
53	291	変わる(*kawaru*, change)*1	369,201	78	453	切る (*kiru*, cut)*1	226,748
54	302	書く (*kaku*, write)*1	344,615	79	456	買う (*kau*,buy)*1	225,436
55	307	よる (*yoru*)*2	342,345	80	458	飲む (*nomu*, drink)*1	224,455

56	310	つく (*tsuku*, attach)*2	340,782	81	459	戻る (*modoru*, return)*1	224,274
57	313	とる (*toru*, take)*2	339,030	82	461	合わせる(*awaseru*, fit)*1	224,028
58	322	頂く (*itadaku*, take)*1	326,225	83	464	いらっしゃる (*irassharu*, come)*1	222,520
59	328	頑張る (*ganbaru*, do one's best)*1	322,837	84	468	残る (*nokoru*, remain)*1	220,486
60	333	始める (*hajimeru*, begin)*1	315,409	85	470	出来る (*dekiru*, can do)*1	219,584
61	335	見せる (*miseru*, show)*1	313,676	86	478	向かう (*mukau*)*1	214,354
62	336	上がる (*agaru*, up)*1	312,565	87	493	降る (*oriru*, get down)*1	208,289
63	355	終わる (*owaru*, finish)*1	294,026	88	494	会う (*au*, meet)*1	208,101
64	359	感じる (*kanjiru*, feel)*1	290,985	89	501	乗る (*noru*, ride)*1	204,365
65	362	帰る (*kaeru*, return)*1	289,341	90	508	調べる (*shiraberu*, search)*1	202,448
66	369	下さる(*kudasaru*, give)*2	282,246	91	514	選ぶ (*erabu*, choose)*1	198,863
67	386	おく (*oku*, put on)*2	268,815	92	520	合う (*au*, fit)*1	196,142
68	388	呼ぶ (*yobu*, call)*1	268,253	93	527	立つ (*tatsu*, stand)*1	194,572
69	389	始まる(*hajimaru*, start)*1	267,644	94	531	置く (*oku*, put on)*1	190,648
70	392	起きる (*okiru*, get up)*1	266,421	95	539	でる(*deru*, get out, exit)*2	187,899
71	393	かかる (*kakaru*)*2	265,463	96	540	描く (*egaku*, draw)*1	187,782
72	400	開く (*hiraku*, open)*1	261,896	97	546	走る (*hashiru*, run)*1	185,448
73	402	決める(*kimeru*, decide)*1	259,893	98	555	なさる (*nasaru*, do)*2	183,528
74	414	話す (*hanasu*, talk)*1	253,373	99	556	伝える (*tsutaeru*, inform)*1	183,445
75	439	やめる (*yameru*, quit)*1	232,015	100	561	すぎる (*sugiru*)*2	181,410

Table 8 has 27 formal verbs and 23 normal verbs and Table 9 has 10 formal verbs and 40 normal verbs. Therefore, the number of formal verbs is larger than the number of normal verbs in the top 50 verbs, although the order is reversed in the following 50 most frequently used verbs. The results show the following features of verbs in the Japanese CCTV corpus: (1) large portions of the uppermost frequently used verbs are formal verbs that are used as functional words written by hiragana letters, and (2) normal verbs frequently appear in the top 100 verbs and this differs from the case of normal nouns in the previous section.

Conclusions

In this paper, we reported the investigation results of the vocabularies in our CCTV corpus. For Japanese language learners, we described the specific details of TV program vocabulary and investigated what kinds of words are necessary for understanding the contents of TV scripts. We used our closed caption TV corpus of over 1 billion words for the investigation of vocabulary. We showed different word statistics from various viewpoints, such as the difference in years and the difference in part-of-speech categories. The results showed the following conclusions:

(1) From the viewpoint of the difference of years, the distribution of the top 100 most frequent words were almost the same during the most recent six years. Therefore, there is a certain tendency for word frequencies in Japanese TV vocabularies.

(2) From the viewpoint of part-of-speech categories, the sum of the frequencies for the top five most frequent part-of-speech categories is over 91% of all tokens. The ratio for the sum of the word frequencies for nouns is about 33.7%, followed in order by particles, verbs, symbols, and auxiliary verbs which are 25.5%, 12.5%, 10.3%, and 9.8%, respectively. On the other hand, nouns and verbs occupy a large amount of vocabularies. The ratios for the types of nouns and verbs are over 94% and 1.4%, respectively, though the ratios of particles, symbols, and auxiliary verbs are only 0.02%, 0.01%, and less than 0.01%, respectively. This indicates that particles, symbols, and auxiliary verbs should be given priority in learning Japanese because they can cover a wide range of the Japanese CCTV corpus by using smaller vocabularies.

(3) From the viewpoint of coverage, the cover ratios for the total frequencies of words against the top 10,000 types of words are 94.8%. Roughly speaking, if Japanese language learners memorize the top 10,000 words on the list, they can obtain the vocabulary necessary to understand over 94.8% of all words used in Japanese TV programs. The list will cover a wide range of TV vocabularies by a limited size of vocabularies, and will become a profitable learning material for Japanese learners.

Acknowledgments

This research was supported by the Grant-in-Aid for Scientific Research (B) (No. 15H02794) and (B) (No. 19H04224) of JSPS.

References

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. Applied Linguistics, 35, 305–327.

Flowerdew, L. (2011). Corpora and Language Education. Palgrave Macmillan.

Newman, J., Baayen, H. & Rice, S. (2011). Corpus-based Studies in Language Use, Language Learning, and Language Documentation. (Language and Computers Studies in Practical Linguistics), Rodopi.

Mochizuki, H. & Shibano, K. (2014). Building Very Large Corpus Containing Useful Rich Materials for Language Learning from Closed Caption TV. World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Volume 2014, No. 1, pp. 1381-1389. Association for the Advancement of Computing in Education (AACE), New Orleans.

Mochizuki, H. & Shibano, K. (2016). Extracting Formulaic Sequences Containing Useful Expressions for Language Learning from Closed Caption TV Corpus, World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, E-Learn 2016, Alexandria, USA, pp. 29-37, November 2016.

Tsujimura, N. (1996). An introduction to Japanese Linguistics. Blackwell Publishers Inc.