EPiC Series in Language and Linguistics

Volume 4, 2019, Pages 33–44

Proceedings of Third Workshop "Computational linguistics and language science"

A Corpus-based Study of Japanese Verb Paradigms (Preliminary results)

Anna Novoselova[1], Alexander Kostyrkin[2]

[1]Russian State University for the Humanities
[2]Institute of Oriental Studies of the Russian Academy of Sciences
annovosyolova@yandex.ru, merukure@yandex.ru

Abstract. The Japanese language has a great variety of verb inflectional suffixes (auxiliaries), each having conjugation of their own. In this paper we propose a corpus-based approach to studying Japanese verb paradigms. Such an approach benefits from identifying possible verb forms on big data of written language. Description of methods and tools used for building databases of verbs and auxiliaries and for parsing verb 7-grams from a Japanese N-gram Corpus is presented.

Keywords: corpus-based study, Japanese language, Japanese dictionary, n-grams corpus, verb paradigms study, verb conjugation.

1 Introduction

1.1 Theoretical Background and Motivation

Japanese verb category, which we treat excluding predicative adjectives, is characterized by a possibility of adding a variety of grammatical elements (in what follows we call them 'auxiliaries'). This variety forms the conjugation paradigm of verbs, and for the Japanese language verb auxiliaries can indicate the following grammatical categories [2]:

1. tense (non-past and past);
2. mood (desiderative, hortative, imperative, indicative, potential, presumptive);
3. voice (causative, passive, etc.);
4. politeness (simple forms, honorific and antihonorific);
5. evidentiality;
6. negation.

Auxiliaries of these categories can be combined in various ways when following a verb stem. Restrictions on possible auxiliary combinations are examined towards their co-occurrence [4], [7], [11] and are considered when describing semantic classes of Japanese verbs. Previous works [1], [2], [10] discern paradigms of active and stative verb classes. For example, the paradigm of active verbs is more extended compared to stative ones at least in expressing mood meanings [2: 76]. Observationally speaking, grammatical features of verbs encode semantic information, and the tendencies of verb endings (sets of auxiliaries) can reveal more information.

It is important to mention that Japanese verb auxiliaries can conjugate. Table 1 illustrates a paradigm of a causative auxiliary *-(sa)seru-*:

Table 1. Paradigms of a Causative Auxiliary -(*sa)seru-*

	(さ)せる *-(sa)seru-*
Imperfective form	(さ)せ *-(sa)se-*
Conjunctive form	(さ)せ *-(sa)se-*
Final form	(さ)せる *-(sa)seru-*
Attributive form	(さ)せる *-(sa)seru-*
Subjunctive form	(さ)せれ *-(sa)sere-*
Imperative form	(さ)せろ *-(sa)sero-*
Imperfective u-connection	(さ)せよ *-(sa)seyo-*

Our research prospective is to try to distinguish verbs not from their meaning but from their paradigm potentialities. In order to do this, it is necessary to elucidate tendencies of verb endings productivity. By way of illustration, according to [2], the passive auxiliary (PASS) (ら)れ *-(ra)re-* cannot appear together with homonymous potential and honorific auxiliaries. However, corpus examples detect that the verb 考える *kangaeru* ‘to think of, to consider’ allows complementarity of passive and potential elements:

(1) kangae-rare-sase-rare-mas-(i)ta
think-POT-CAUS-PASS-ADR-PST
‘made someone able to think’

A corpus-based method makes it possible to retrieve real language data and to detect lacunae in the verb paradigms [14]. Taking into consideration the fact that grammatical behavior of verbs is influenced by its semantics, this study is seen as a preliminary but necessary step in relation to analyzing the semantics of verb endings and the verbs themselves. This can help us to build more complete paradigms than provided by grammars or automatic generators, which cannot foresee any lacuna or biases. Additionally, the results of this study are a reliable dataset of language use with frequency parameters, which eventually enables scientists to conduct research of the Japanese verb system. These results are going to be used for further semantic classification of Japanese verbs. Moreover, it might help to classify newly-coined verbs.

The final products of the study are presented in the form of SQLite databases.

1.2 Material

Our experiment was conducted on a corpus by Yata Susuno, “Nihongo Web Corpus 2010 - N-gram Corpus”. This corpus gives access to data from 2-grams to 7-grams. A morphological tag set is also available. Tokens were included into this corpus only if they appear in the texts more than 10 times. By a token the authors mean a morpheme. For our study we used the corpus of 7-grams. There are 347,097,023 unique 7-grams in this corpus.

Information for parsing the data from the Web corpus was taken from a Japanese dictionary, IPADIC (version 2.7.0). This dictionary is based on a version of an IPA (Information-technology Promotion Agency) Part of Speech tag set. There are 233,634 words, which correspond to 'morphemes' of the corpus data. Besides, there are manuals and files providing part of speech and grammar information. The method of tagging the information in the dictionary is not entirely convenient for our study. According to the traditional Japanese part of speech classification, auxiliaries are characterized as auxiliary verbs (助動詞 *jodōshi*). For this reason, all verb auxiliaries, auxiliary verbs and the latter part of compound verbs (such as ある *aru* 'to be', 御座る *gozaru* 'to be, to exist', 無く *naku* 'to be absent', がる *garu* 'to be felt as') are retained in Auxil.dic file. We took this into consideration when compiling data for the research.

There are several reasons for using the IPADIC dictionary as a provider of information for the parser in our study. Firstly, the corpus we have used is based on it: the n-grams are composed of elements according to the dictionary entries. Secondly, the terminology accepted by the authors of the dictionary is used in most of the Japanese NLP projects. Further, all the posterior dictionaries are modifications of this particular dictionary.

1.3 Research Limitations and Terminology Comments

A morpheme in written Japanese can be represented by ideographic, katakana or hiragana writing. We did not approach the tasks of unification of verb written representation but restricted the input data. While collecting the data we limited our research only to verbs with ideographic character. Dealing with auxiliaries, the process was opposite: first we collected all the auxiliary forms, and then we reduced their dictionary forms to hiragana written ones.

In order to distinguish parts (morphs) of a verb, we are using the term 'base' as a bare form without any final element such as an auxiliary or a vowel of a stem. For instance, the base form for 取ります *tor-imasu* 'someone takes' is *tor-*. The problem of Japanese morpheme boundaries is well-studied and is described in particular in [1], [5].

What we call an infinitive form is the dictionary form of verbs [3].

We accept the automatic way of dividing the words into morphemes performed by the corpus developers. We keep in mind that multiple repetitions of auxiliaries might be emphasizing or emotional and refer to some extra-semantic level. Still, as they appear in the corpus, we examine them as well. A more questionable example is illustrated in (2): 死ねねねねねね, a form of 死ぬ *sinu* 'to die' can be an imperative base of a verb (*sine*), where there is a repetition of the last syllable of a verb form, which is homonymous to an emphatic particle ね *ne*:

(2) sin.e-ne-ne-ne-ne-ne
die.IMP-$^{?}$PRT-$^{?}$PRT-$^{?}$PRT-$^{?}$PRT-$^{?}$PRT
'Die! Shall you?'

One of the difficulties arising from allowing such forms is that when we filter language material by formal features (such as parts of speech or auxiliary class) there is a possibility to get a huge number of erroneous examples. Such endings are not representing real paradigms and should be distinguished from them. In such and ending as なしなしなしありなし (3) there are no verbs presented, but the automatic parser gave it as output as a combination of auxiliaries, which is grammatically incorrect and must be deleted from the data. なし nasi is a finite form which adds only to the imperfective base and cannot be followed by any form of the verb aru ‘to be; to exist’ in its conjunctive base.

(3) nasi-nasi-nasi-ari-nasi
NEG-NEG-NEG-AUX-NEG
‘no-no-no-exits-no‘

We are not providing description of Japanese verb conjugation patterns here and refer to [9], [4] for grammatical theory.

Speaking about the semantic classes of the verbs we accord with the terminology accepted in [10]. The most common differentiation is between active and stative verbs.

Active verbs have additional characteristics of durativity, telicity and volition. Our preliminary step towards verb paradigms study is conducted on the material of Balanced Corpus of Contemporary Written Japanese (BCCWJ) in order to test hypothetical verb forms for instances of different action and stative verbs in Active Voice. A concise distribution of auxiliaries in Table 2 is presented on the grounds of BCCWJ study. All of the action verbs presented in the table are semantically telic. Categories included in Table 2 are: present tense (PRS), politeness (ADR for ‘addressive’), durativity (PRG for ‘progressive’), resultative (RES), purpose (PURP), and converb (CNV). Grey fields in the table indicate forms not found in the corpus and therefore viewed as unlikely to be produced.

Table 2. Examples of Semantically Different Verb Paradigms.

	Action			Stative		
	食べる *taberu* ‘to eat’	折る *oru* ‘to break’	眠る *nemuru* ‘to sleep’	似る *niru* ‘to resemble’	聳える *sobieru* ‘to tower’	表す *arawasu* ‘to indicate’
	Durative Volitional	Non-durative Volitional	Durative Non-volitional			
RES-PRS (+ aru)	食べてある *tabe-te aru*	折ってある *ot-te aru*				表してある *arawas-ite aru*
RES-ADR-PRS (+ aru)		折ってあります *ot-te ar-imas.u*	眠ってあります *nemut-te ar-imas.u*			表してあります *arawas-ite ar-imas.u*
PRG2-PRS		折りつつある *or-itsutsu aru*	眠りつつある *nemur-itsutsu aru*	似つつある *ni-tsutsu aru*	聳えつつある *sobie-tsutsu aru*	表しつつある *arawas-itsutsu aru*
PRG2-ADR-PRS		折りつつあります	眠りつつあります	似つつあります	聳えつつあります	表しつつあります

		or-itsutsu arias'	*nemur-itsutsu ar-imas.u*	*ni-tsutsu ar-imas.u*	*sobie-tsu-tsu ar-imas.u*	*arawas-itsutsu ar-imas.u*
PURP. PRS	食べてみる *tabe-te-mir.u*	折ってみる *ot-te-mir.u*	眠ってみる *nemut-te-mir.u*	似てみる *ni-te-mir.u*		表してみる *arawas-ite-mir.u*
PURP-ADR. PRS	食べてみます *tabe-te-mi-mas.u*		眠ってみます *nemut-te-mi-mas.u*	似てみます *ni-te-mi-mas.u*		表してみます *arawas-ite-mi-mas.u*

The extraction here lets us see that among both the classes there can be verbs that act differently. It is seen in Table 2 that the non-durative verb 折る *oru* ‘to break’ allows the same number of forms as the durative 眠る *nemuru* ‘to sleep’ does, but the missing forms are different. At the same time the paradigm of the durative and volitional verb 食べる *taberu* ‘to eat’ has little in common with the paradigms of either durative 眠る *nemuru* ‘to sleep’ or volitional 折る *oru* ‘to break’. Action verbs have more characteristics that stative verbs, still stative verbs do not tend to have similar paradigms – 似る *niru* ‘to resemble’ and 聳える *sobieru* ‘to tower’ are missing resultative forms, but 聳える *sobieru* ‘to tower’ has no form of purpose. On the other hand, in Table 2 the verb 表す*arawasu* ‘to indicate’ has no forms missing.

An assumption can be made that there might be some meaning that allows to distinguish verbs of these two classes explicitly – and this leads to the task of learning all verb potentialities and clustering them automatically on the base of their grammatical features.

2 Data Processing

2.1 Data Extraction and Databases Creation

To parse the corpus data, we used the dictionary files of IPADIC, which contain lists of POS and provide grammar and phonetic information. These files are stored in a dictionary format with brackets separated string entries (LISP Processing language). For our tasks we transformed these files to SQLite database and converted the data from the original EUC-JP encoding into UTF-8 encoding.

Each line of the dictionary files begins with a base form of a word as its heading. Information in the line characterizes not the word itself (a lexeme) but the form representing it. For instance, the verb file (Verb.dic) consists of 130,750 entries, which have verb bases or verbs stems as their key. Example (5) demonstrates one entry of 謝ら *ayamara-* ‘to apologize’ in its imperfective form. Details and explanation of the information held in such a line are presented in table 4. Files with other parts of speech are constructed identically.

(4) 謝ら,780,780,7130, 動詞,自立,*,*,五段・ラ行,未然形, 謝る,アヤマラ,アヤマラ

Of all the dictionary files we handled files with verbs, auxiliary verbs and postpositional affixes (these two files include verbal inflectional auxiliaries), and a file with particles. The last file was used because a verb ending can have a final particle.

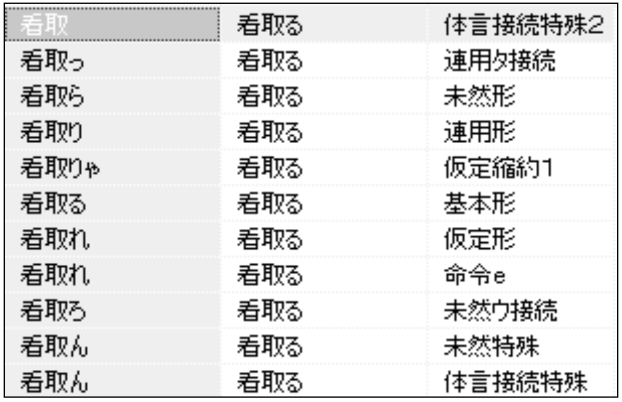
Comments to the information given and examples of the verb 賭する *tosuru* ‘to bet’ and its dictionary form 賭す *tosu*, an addressive auxiliary ます *masu* and a converb particleて *te* entries are provided below in Table 3.

Table 3. Examples of Information in the Dictionary Files

	Verb Entry Examples		Auxiliary Verb Entry Example	Postposition Entry Example
Text form	賭する (*tosuru*)	賭す (*tosu*)	ます (*masu*)	て (*te*)
Morpheme occurrence cost based on morpheme occurrence probability	<…> 7150	<…> 6837	<…> 5537	<…> 5170
POS name	動詞 (*‘verb’*)	動詞 (*‘verb’*)	助動詞 (’*auxiliary verb*’)	助詞 (*‘verb’*)
POS category	自立 (*‘main verb’*)	自立 (*‘main verb’*)	*	接続助詞 (*‘particle conjunctive’*)
Inflection type information	*,*	*,*	*,*	*,*
Inflection type	サ変・−スル (’*verbs formed with “suru”*’)	サ変・−スル (’*verbs formed with “suru”*’)	特殊・マス (*‘special “masu” inflection’*)	*
Written basic form	基本形 (*‘basic form’*)	文語基本形 (*classical base form)*	基本形 (*‘basic form’*)	*
Dictionary form	賭する(*tosuru*)	賭する(*tosuru*)	ます (*masu*)	て (*te*)
Reading in katakana characters	トスル (*tesuru*)	トス (*tosu*)	マス(*masu*)	テ (*te*)
Pronunciation (written in katakana)	トスル(*tesuru*)	トス (*tosu*)	マス(*masu*)	テ(*te*)

The files give information about a base form of a lexeme. Even if it is an affix, there still can be several representations of it (such an example is provided above for a causative auxiliary in Table 1). Since all the auxiliaries attach to a certain verb base, our next step was to join all the entries of the same verbs. For example, there are 11 entries for the verb 看取る *mitor-u* ‘to attend someone’s deathbed' which are verb bases of this verb. All of them were added to the verb data base. Table 4 shows how these bases are kept in the Verbs database ('base type' column coincides the 'inflection type' row from Table 4). For reference on the base (inflection) type terminology see [3].

Table 4. Base Forms of 看取る *mitor-u* 'to attend someone's deathbed'



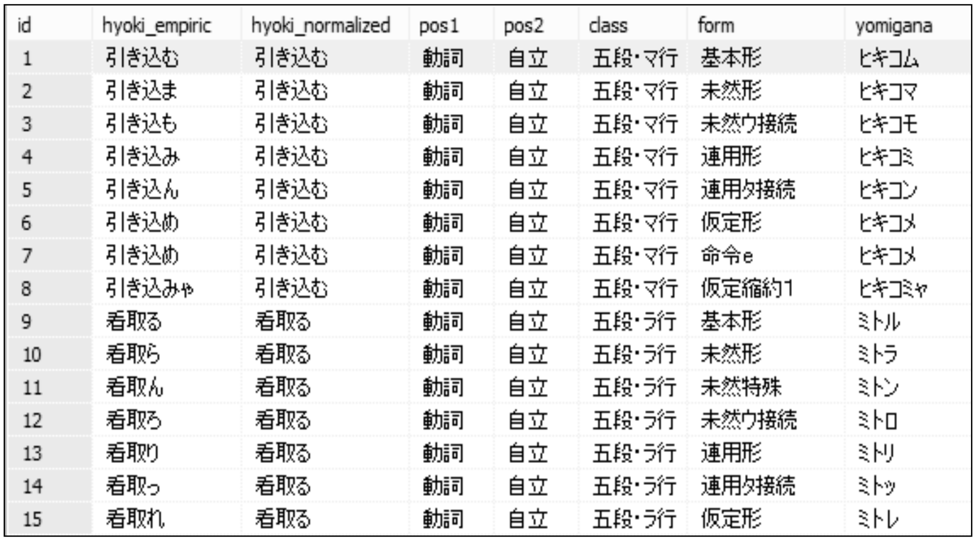
Verb base	Dictionary form	Inflection type
看取	看取る	体言接続特殊2
看取っ	看取る	連用タ接続
看取ら	看取る	未然形
看取り	看取る	連用形
看取りゃ	看取る	仮定縮約1
看取る	看取る	基本形
看取れ	看取る	仮定形
看取れ	看取る	命令e
看取ろ	看取る	未然ウ接続
看取ん	看取る	未然特殊
看取ん	看取る	体言接続特殊

Grammar information (POS, inflection type and inflection type information) is given to the extent of ambiguous forms of a verb. For the same reason we preserve the phonetic information (*yomigana* column) as some ideographic characters can perform different readings and meaning and this helps to disambiguate different verbs of the same form.

The verb data base generated from the dictionary material consists of 79,639 empiric verb forms which count for 8,858 individual verbs. The first 15 entries of the verb database are presented in Table 5. Three verbs – 有る *aru* 'to be' (about inanimate subjects mainly), いる *iru* 'to be; to exist' (about animate subjects) and くる *kuru* 'to come' were excluded from the group of verbs because of their high frequency and mostly auxiliary usage. 有る *aru* and いる *iru* can be found in our generated database of auxiliaries – Aux_verbs Database. The name for this database is following the terminology of IPADIC authors, which view all verb inflectional elements as auxiliary verbs. The auxiliary database is similar in its structure to the verb database except the yomigana column.

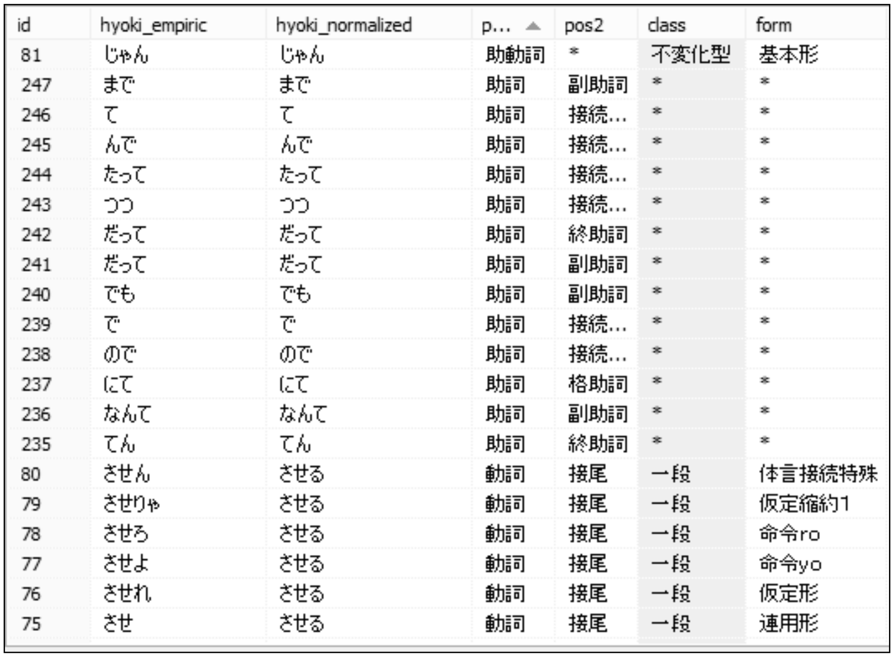
Overall there are 247 auxiliaries in the database which correspond to 42 individual dictionary forms of auxiliaries. A sample of auxiliary database is presented in Table 6

Table 5. Sample of 15 Entries from the Verb Database



id	hyoki_empiric	hyoki_normalized	pos1	pos2	class	form	yomigana
1	引き込む	引き込む	動詞	自立	五段・マ行	基本形	ヒキコム
2	引き込ま	引き込む	動詞	自立	五段・マ行	未然形	ヒキコマ
3	引き込も	引き込む	動詞	自立	五段・マ行	未然ウ接続	ヒキコモ
4	引き込み	引き込む	動詞	自立	五段・マ行	連用形	ヒキコミ
5	引き込ん	引き込む	動詞	自立	五段・マ行	連用タ接続	ヒキコン
6	引き込め	引き込む	動詞	自立	五段・マ行	仮定形	ヒキコメ
7	引き込め	引き込む	動詞	自立	五段・マ行	命令e	ヒキコメ
8	引き込みゃ	引き込む	動詞	自立	五段・マ行	仮定縮約1	ヒキコミャ
9	看取る	看取る	動詞	自立	五段・ラ行	基本形	ミトル
10	看取ら	看取る	動詞	自立	五段・ラ行	未然形	ミトラ
11	看取ん	看取る	動詞	自立	五段・ラ行	未然特殊	ミトン
12	看取ろ	看取る	動詞	自立	五段・ラ行	未然ウ接続	ミトロ
13	看取り	看取る	動詞	自立	五段・ラ行	連用形	ミトリ
14	看取っ	看取る	動詞	自立	五段・ラ行	連用タ接続	ミトッ
15	看取れ	看取る	動詞	自立	五段・ラ行	仮定形	ミトレ

Table 6. Sample of 15 entries from the Aux_Verb Database



id	hyoki_empiric	hyoki_normalized	p... ▲	pos2	class	form
81	じゃん	じゃん	助動詞	*	不変化型	基本形
247	まで	まで	助詞	副助詞	*	*
246	て	て	助詞	接続...	*	*
245	んで	んで	助詞	接続...	*	*
244	たって	たって	助詞	接続...	*	*
243	つつ	つつ	助詞	接続...	*	*
242	だって	だって	助詞	終助詞	*	*
241	だって	だって	助詞	副助詞	*	*
240	でも	でも	助詞	副助詞	*	*
239	で	で	助詞	接続...	*	*
238	ので	ので	助詞	接続...	*	*
237	にて	にて	助詞	格助詞	*	*
236	なんて	なんて	助詞	副助詞	*	*
235	てん	てん	助詞	終助詞	*	*
80	させん	させる	動詞	接尾	一段	体言接続特殊
79	させりゃ	させる	動詞	接尾	一段	仮定縮約1
78	させろ	させる	動詞	接尾	一段	命令ro
77	させよ	させる	動詞	接尾	一段	命令yo
76	させれ	させる	動詞	接尾	一段	仮定形
75	させ	させる	動詞	接尾	一段	連用形

2.2 Parsing the corpus data

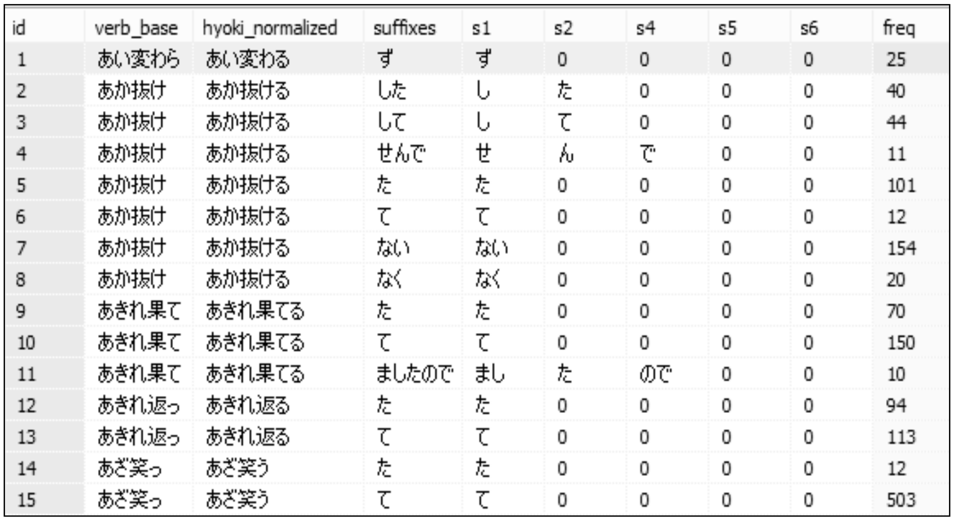
The parsing algorithm for the 7-gram corpus data and its preprocessing steps are as follows. First the string part was checked for being a verb base. If it was, then the tail of the string was remembered and if parts of the tail were in the list of auxiliaries, the ending was written to an output file. If there were already existing verb bases, the endings were added to it as their value and the number of occurrences was added.

A simplified algorithm can be seen as a three-stepped one:

- create a verb dictionary from the verb database:
  {empiric verb form: index of its dictionary form};
- create an auxiliary dictionary from the auxiliary database:
  {empiric auxiliary form: index of its dictionary form};
- if the first element of an n-gram is in the verb dictionary keys:
  - check the following elements for being in the auxiliary dictionary keys;
  - remember indices for each element found.

As the POS and Inflection information can be found in the databases, the final output file (which is the database of 7-grams) consists only of a dictionary verb form plus its index, a list of endings with indices for each of them and frequency of the verb form. Indices enable quick turning to related databases and finding information for every constituent of an n-gram.

Table 7. Sample of N-grams Database entries



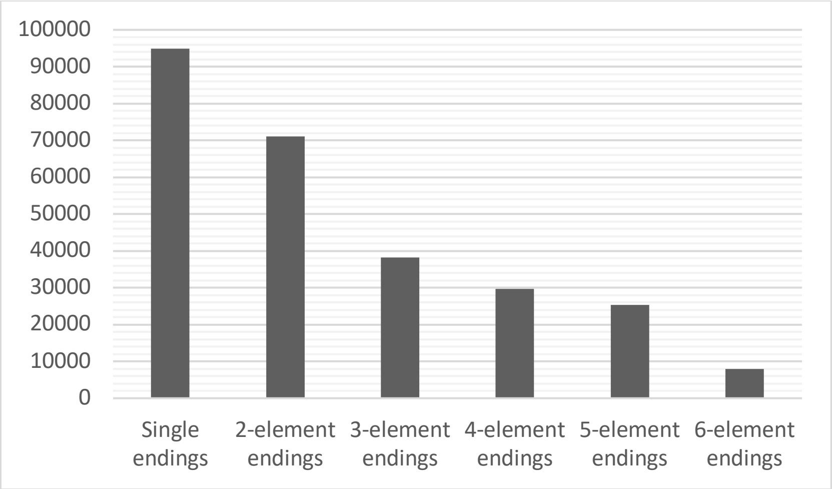
id	verb_base	hyoki_normalized	suffixes	s1	s2	s4	s5	s6	freq
1	あい変わら	あい変わる	ず	ず	0	0	0	0	25
2	あか抜け	あか抜ける	した	し	た	0	0	0	40
3	あか抜け	あか抜ける	して	し	て	0	0	0	44
4	あか抜け	あか抜ける	せんで	せ	ん	で	0	0	11
5	あか抜け	あか抜ける	た	た	0	0	0	0	101
6	あか抜け	あか抜ける	て	て	0	0	0	0	12
7	あか抜け	あか抜ける	ない	ない	0	0	0	0	154
8	あか抜け	あか抜ける	なく	なく	0	0	0	0	20
9	あきれ果て	あきれ果てる	た	た	0	0	0	0	70
10	あきれ果て	あきれ果てる	て	て	0	0	0	0	150
11	あきれ果て	あきれ果てる	ましたので	まし	た	ので	0	0	10
12	あきれ返っ	あきれ返る	た	た	0	0	0	0	94
13	あきれ返っ	あきれ返る	て	て	0	0	0	0	113
14	あざ笑っ	あざ笑う	た	た	0	0	0	0	12
15	あざ笑っ	あざ笑う	て	て	0	0	0	0	503

3 Results and Analysis

The output file of verb n-grams contains 267,024 verb forms (a base plus an ending). All these verbforms correspond to 6,830 verbs with an average of about 40 possible suffix endings per verb.

The frequency of verb endings depending on their length is shown in Picture 1. The fact that the most frequent endings have only one auxiliary (94,880) is neither non-standard nor unexpected. They count for almost one third of all the endings.

Fig. 1. Frequency Characteristics of Endings Depending on the Number of their Auxiliaries



It is reasonable that the most frequent endings consist of one and two auxiliaries. The grammatical functions of these auxiliaries are shown in Table 8. Number of entries is frequency according to the number of verbs that can be used with corresponding endings.

Table 8. The Most Frequent Verb Auxiliaries

№	Auxiliary	Comment	Number of entries
1.	て *te*	converb	13,317
2.	た *ta*	past tense	9,712
3.	ない *nai*	negation	6,338
4.	で *de*	converb	4,827
5.	れ *re*	passive/protentional/honorific	3,809
6.	う *u*	‘auxiliary verb’ from Japanese linguistic tradition. We assume it to be presumptive auxiliary	3,508
7.	し *shi*	Might be an error from する *suru* ‘to do’, whose form し *shi* is added to nominalized verbs	2,920
8.	ず *zu*	converb plus negation	2,850
9.	れる *reru*	passive/protentional/honorific	2,745
10.	たい *tai*	desiderative	2,694
11.	れ-た *re-ta*	passive/protentional/honorific plus past tense	2,610
12.	ます *masu*	addressive, present tense	2,563

13.	れ-て *re-te*	passive/protentional/honorific plus converb	2,521
14.	まし-た *mashi-ta*	addressive, past tense	2,343
15.	な-かった *na-katta*	negation, past tense	2,273

A more interesting case is six-element-endings. Ten most frequent ones are shown in Table 9. It is noticeable that all of the forms have a negation auxiliary, so we can assume that the language allows large auxiliary stacking when there is a negation and even allows repetition of them in different positions. As the negation can relate to any meaning provided by other auxiliaries in the ending, a study of possible positions for a negation auxiliary and negation itself would be valuable. Table 9 shows ten most frequently used endings of six elements (presumptive is glossed as PMT, SBST is substantiviser, past tense is PST, copula is COP, particle is PRT, negative is NEG, hortative is HOR). う *u* is not glossed separately but in the corpus the character is viewed as an individual gram.

Table 9. Ten Most Frequent Six-length Endings

Verb six-length ending	Glossing	Romaji representation	Number of entries
てたんだろうな	CNV-PST-SBST-PMT-PRT	te-ta-n-darou-na	222
たんじゃないでしょう	PST-SBST-COP-NEG-PMT	ta-n-ja-nai-desyou	208
たんじゃないだろう	PST-SBST-COP-NEG-PMT	ta-n-ja-nai-darou	130
Verb six-length ending	**Glossing**	**Romaji representation**	**Number of entries**
てたんでしょうね	CNV-PST-CNV-HOR-PRT	te-ta-n-desyou-ne	128
ないんじゃないでしょう	NEG- SBST-COP-NEG-CNV-HOR	nai-n-ja-nai-desyou	123
なかったんだろうな	NEG-PST- SBST-PMT-PRT	na-kattan-da-rou-na	122
てないんだろうな	CNV-NEG- SBST-PMT-PRT	te-nai-n-darou-na	115
てたんだろうね	CNV-PST- SBST-PMT-PRT	te-ta-n-darou-ne	99
てませんでしたね	CNV-ADR-NEG-COP-PST-PRT	te-mase-n-desi-ta-ne	96
てなかったんですね	CNV-NEG-PST-NEG-COP.ADR-PRT	te-na-katta-n-desu-ne	92

4 Conclusions

In this work we have described principles of analyzing Japanese corpus material from the prospective of examining verb paradigms. The study is based on 7-grams corpus material and our work with

Balanced Corpus of Contemporary Written Japanese (BCCWJ). The latter step is seen as a preliminary stage in which we checked some hypotheses for grammar collocability of verbs, provided corpus evidence for accepting or rejecting our prospects and established some preliminary findings. In future we intend to make a comparative study of the results from BCCWJ and the results of parsing the 7-gram corpus.

In order to parse the corpus data, it was necessary to investigate the ideology of the dictionary and the tag set it is based on. During this process it was possible to reveal some inaccuracies that lead to erroneous parsing.

Even at this stage we have an informative material which allows to view the paradigms of a Japanese verbs using modern methods. For example, combinations of grammar elements which have been restricted by authors of existing grammars or which have not been considered of are found.

The data extracted still needs normalizing and preprocessing tools as filters that were created on the base of grammar and language intuition are not enough. There are combinations which have nothing to do with verb but because of some erroneous parsing they appear and are quite frequent. Normalizing and cleaning the defective data are some of further steps.

The research described is one of the steps of our Japanese verb semantic classification project. The aim of the study was to extract information representing verb and auxiliary collaboration from real language data. Different special program scripts were created for corpora parsing and retrieving information from dictionary resources. The results provide profound data of linguistical and statistical characteristics.

References

1. Alpatov V.: Sematic Classes of Japanese Verbs. [semanticheskie klassy yaponskih glagolov] (in Russian) In: Aktual'nye voprosy yaponskogo i obschego yazykoznaniya, pp. 45-57. (2005)
2. Alpatov V., Arkadiev P., Podlesskaya V.: Theoretical Grammar of the Japanese language. [teoreticheskaya grammatika yaponskogo yazyka] (in Russian), Vol.1 Moscow (2008)
3. Asahara, M., Matsumoto, Y.: Masayuki Asahara and Yuji Matsumoto, IPADIC Version 2.7.0 User's Manual. Nara Institute of Science and Technology (2003)
4. Chino N.: Japanese Verbs at a Glance. Kodansha International (2001)
5. Kamiya E.: Review of auxiliary verbs co-occurrence. (in Japanese) In: Gobun, Vol. 40, pp. 21:35. Osaka University (1982)
6. Kieda M.: Grammar of the Japanese Language. [grammatika yaponskogo yazyka] (in Russian), Vol. 1, pp. 127-206; 344-507. Moscow (2002)
7. Kitahara Y.: Structural Review of auxiliary verbs co-occurrence. (in Japanese) Kokunogaku (1970)
8. Lampkin, R. L.: Japanese Verbs & Essentials of Grammar: a Practical Guide to the Mastery of Japanese. New York (2004)
9. Lavrentiev B.: Practical Grammar of the Japanese Language. [practicheskaya grammatika yaponskogo yazyka] (in Russian) pp. 88-224. Moscow (2002)
10. Martin, S. E.: A Reference Grammar of Japanese. Tokyo (1991)
11. Mizutani S.: – Experiment of Examining the system of 'go' co-occurrence. Kotoba no Kenkyu (1959)
12. Yasuharu D., Nakamura J., Ogiso T., Ogura H.: A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In: Proceeding of the 6th Language Resources and Evaluation Conference (LREC 2008), pp. 1019–1024 (2008)
13. Tadoharu T.: Corpus-based Grammar Study. In: Nihongogaku, Vol.22, pp. 174-185 (2003)
14. Balanced Corpus of Contemporary Written Japanese (BCCWJ) http://www.kotonoha.gr.jp/shonagon