Sampling of Word Forms and Compound Words Based on N-gram Statistics (Analysis of Japanese language in the Heian Period)

Miyuki Kondo

(A Method for Studying Waka from the Dynastic Period , Kasama-shoin, 2015, Chapter 12)

1. Introduction

There are many points of view and perspectives on how to establish criteria for identifying an individual word in classical Japanese or modern Japanese. In addition, extensive research has been conducted on this process, which has produced no consensus in regard to the units that should be used to identify a word. For example, in several studies conducted by the *National Language Research Institute*, two types of units have been proposed: alpha units and beta units [1], which reveal the difficulty of recognizing words in Japanese. One extreme example is the case of compound words. To date, there have been four approaches to defining and analyzing compound words: 1) the morphological approaches [2]; 2) phonological approaches [3]; 3) statistical approaches [4]; and 4) semantic approaches [5]. As there have been numerous comparisons with other foreign languages [6], it may appear as if this topic has been discussed from every possible angle. However, there are problems that have not necessarily been resolved such as the vague methods in which compound words are identified as items in indices and dictionaries [7] as well as cases in which the boundaries with words, customary expressions, and texts are unclear [8]. One of the reasons for such difficulty is the meshing of theory and reality, although it is probably related to the fact that identification has been primarily based on the researchers' introspective opinions. There are more than a few instances in which discrepancies in the selection of compound words as items in indices and dictionaries can be described as reflections of differences in the editors' subjective opinions and linguistic theories. Of course, one cannot doubt the importance of introspection in research about grammar, vocabulary, and diction. However, when classical Japanese (not modern Japanese) is the subject of attempts to sample and analyze the forms and semantic features of compound words, a more appropriate grasp of linguistic phenomena is required in addition to introspection. Therefore, this study proposes a method of sampling and analyzing items from a corpus of compound words by using statistical processing techniques based on literary texts from the Heian Period. In addition, this study demonstrates one aspect of the knowledge gained from this approach.

2. Linguistic Structure of Japanese and the Rules of Regular Grammar

This paper points out an aspect that has gone almost unnoticed in previous research on compound words: the characteristics of Japanese linguistic structures based on statistical processing of a corpus, especially in regard to Japanese words and structures. Regular grammar, also known as "type 3 grammar," is a category of formal grammar. Unlike phrase structure grammar (type 2 grammar) in which widely separated elements act in accordance with one another, the grammatical rules of regular grammar generally involve a simple consecutive chain of elements. Regarding the rules of regular grammar, the following two aspects are inherent in Japanese linguistic structures.

First, there are the cases in which two independent words are joined together to create a compound such as *hashiridasu* (*hashiru* “run” + *dasu* “send out”) “break into a run,” *mejirushi* (*me* “eye” + *shirushi* “symbol”) “distinguishing mark,” or *suzukuri* (*su* “nest” + *tsukuru* “make”) “building a nest.” In addition, there are other cases in which an independent word is followed by a case particle or suffix, as seen in *na no hana* (*na* “rape plant” + *no* possessive particle + *hana* “flower”) “rape blossom,” *nikumarekko* (*nikumareru* “be hated” + *ko* “child”) “unpopular child,” or *torarezon* (*torareru* “be taken” + *zon* “loss”) “loss due to theft or cheating.” Some resemblance to textual grammatical rules can be seen in these compound rules, and as Okitsu Keiichiro et al. [9] had pointed out, these are essentially based on phrase structure rules (type 2 grammar). In fact, one example in which the elements do not perfectly line up in order from the beginning (a right-branching structure in a tree diagram of phrase structure rules) is *ooyamazakura* (*oo* “big” ((*yama* “mountain”) (*sakura* “cherry tree”)). However, examples like these are exceptions, which are often used for the names of plants and animals. Basically, the majority of compounds are either cases in which a predicate is attached to a complement (*yamanobori*: *yama* “mountain” + *nobori* “climbing” = “mountain climbing” or *ootonogomoru*: *ootono* “great hall” + *komoru* “be shut in” = “sleep, rest” [honorific form]) or cases that follow the rules of regular grammar in which ancillary forms are attached to an independent word (*umarenagara*: *umareru* “be born” + *nagara* “in the course of” = “inborn, congenital” or *nantoshitemo nani* “what” + *to* “with” + *shite* “doing” + *mo* “even, also” = “no matter what”). Mizutani Shizuo [10] offered a theory of type 3 grammar (regular grammar) in which compounding was framed as a problem of the characteristics of Japanese conjugations and grammar. Regarding this theory, Shizuo stated, “Japanese conjugational forms are determined by what kinds of linguistic elements immediately follow them, barring any restrictions on agreement. In other words, if a certain conjugational form appears, the range of linguistic elements that can appear immediately after it is determined.” This remark could be conceivably extended to linguistic structures.

Second, there is the problem of creating new words by combining existing *kanji* (Chinese-derived characters) as seen in *hanamichi* (*hana* “flower” + *michi* “road”) “a bridge-like stage approach in kabuki” or *Nihonsei* (*Nihon* “Japan” + *sei* “manufacturing”) “made in Japan.” It is believed that the logographic (formal element) nature of the symbols allows them to be combined to form new words. Typical examples of these so-called “temporary words” include *taiBeikoushou* (*tai* “facing” + *Bei* “the United States” [an abbreviation of *Beikoku*] + *koushou* “negotiations”) “negotiations with the United States” and *seidenryokukadoryoku (sei* “reduce” + *denryoku* “electrical power” + *ka* “change” + *doryoku* “effort”) “efforts to switch over to the use of less electrical power.” However, these are additional examples where there are simple regular grammar rules in which elements appear if a certain *kanji* appears. This shows that these particular rules play a major role in the formation of compounds.

As seen above, we must first be aware of the rules of regular grammar (type 3 grammar) before creating a theoretical basis for the formation of compound words. However, even if studies confined to case particles and suffixes are included, there has been minimal progress in research since Mizutani's aforementioned study. One factor that has prevented further development of Mizutani's work is the limitations of previous tools and data during the handling of character strings and regular structures, which requires a statistical approach and a certain amount of statistical processing with computers.

However, recent technological advancements in computer speed and capacity have enabled employing even larger memory devices than ever before. In addition, with the restrictions in those areas eliminated, new groundbreaking techniques have been devised that promise more developments when analyzing the regular structures of Japanese linguistic components. One example is the study by Nagao Makoto & Mori Shinsuke in 1993 titled "Dai kibo Nihongo tekisuto no n-gram toueki no tsukurikata to goku no jidou chuushitsu" ("Creation of N-gram Statistics for Large-Scale Japanese Texts and Automatic Sampling of Phrases"). The following section includes an overview of this technique as well as a discussion on its application of analyzing classical Japanese.

3. Identifying the Word Forms of Classical Japanese with Information Theory N-Gram Statistics

Developed by Claude Shannon, n-gram statistics is one of the basic theories of information science [12]. However, given its characteristic as a theory that analyzes linguistic information as the probability of series of units, it is also very appropriate for dealing with the linguistic structures of languages such as Japanese. The technique developed by Nagao & Mori (1993) may be called a practical application of Shannon's n-gram analysis to linguistics. It utilizes independent algorithms to enable extremely high-speed sampling and provide a statistical analysis of chains or strings of characters. Simply stated, the principle behind it is to exhaustively extract strings of two characters, three characters, four characters, … *n* characters (where *n* is an arbitrary number) from a text. Because this process enables finding all connection patterns of the characters, it samples not only combinations of characters but the frequency at which word forms occur. In addition to compound words such as the aforementioned *suzukuri* ("nesting") and *taiBeikoushou* ("negotiations with the United States"), this technique is inherently able to select lengthy strings such as *Nihongo kaiseki shisutemu kaihatsu kenkyuujo shunin kenkyuuin* ("Japanese language analytical system development institute principal researcher") In particular, Nagao & Mori's research has analyzed large-scale corpuses from electronic publications (29.36 million characters, 59 MB). They have succeeded in exhaustively and automatically sampling the types of forms that previous research has referred to as "compound forms" such as *shinakereba naranai* (*shinakereba* "if [someone] does not do" + *naranai* "will not become") "must do" (Nagao et al. refer to these as "compound character strings"). In addition, their research sampled groups of words and phrases with a strong tendency to occur together (collocations and idioms), such as *eikyou wo* "influence + (direct object particle)" with *ataeru* "give (abstract)" or *ukeru* "receive, undergo" to produce *eikyou wo ataeru* "exert an influence" and *eikyou wo ukeru* "be influenced."

We can expect further developments from researching the word forms and linguistic structures of contemporary Japanese. The above-mentioned techniques are also effective for analyzing corpuses in classical Japanese. This study conducts joint research by applying n-gram analysis to classical Japanese and classical literature as well as using Nagao & Mori's program [13]. However, we are not simply utilizing classical Japanese as our corpus. Instead, we are moving forward with two new developments: (1) introducing an approach of sampling word forms among texts with phase differences, and (2) constructing a system for comparing overall character strings among multiple texts by using the n-gram set operator method. In this paper, "phase differences" means gender differences, differences among works such as *Genji monogatari* and *Utsubo monogatari*, and genre differences between prose works (*Genji monogatari*) and poetical works (*Kokinwakashuu*). By comparing overall character strings between texts with phase differences, we can extract all word forms common between the compared texts as well as those that are unique to a particular text. Next, we compare all character strings in the poems written in the *Kokinwakashuu* beginning with the poems created by the male poets.

A. Words, Compounds, and Suffixes

1. Nouns and Compound Nouns:
a. Plants:
ominaeshi "valerian," *momijiba* "colored autumn leaves," *ume no hana* "plum blossom," *hagi* "bush clover," *fujibakama* "thoroughwort," *wakana* "young greens," *wasuregusa* "day lily."
b. Insects:
utsusemi "cicada."
c. Celestial phenomena:
amanogawa "the Milky Way," *tanabata* "the Star Festival," *kumoi* "cloudy place," *fuku kaze* "blowing wind."
d. Colors:
shirayuki "white snow," *shirakumo* "white cloud," *nishiki* "brocade," *shiratama* "white jade," *shiratsuyu* "white dew."
e. Miscellaneous:
aki no no "autumn meadow," *haru no yamabe* "mountainous area in the spring."

2. Verbs:
nabiku "trail off," *wataramu* "cross" (tentative form).

3. Adjectives:
tsurenaki "heartless," *samuku* "cold" (adverbial form), *samumi* "being cold," *yow o samumi* "because the night is cold."

4. Agglutinations of ancillary forms

-karikeru (conjunctive adjectival form + subjectivizing suffix [attributive]), -karikeri (conjunctive adjectival form + subjectivizing suffix [predicative]), -karikere (conjunctive adjectival form + subjectivizing suffix [conjunctive]), -karu beki (conjunctive adjectival from + beki “ought to”), naranaku ni (naru copula + naku “not” + ni “to”) “although [it] is not,” arikere (aru “be” + keri “subjectivizing suffix” [conjunctive form]), aru kana (aru “be” + kana reflective exclamation) “I wonder if there is,” shinakereba (suru “do” + nai “not” + keri “conjunctive adjectival form” + ba “if, when”) “when [someone] does not do,” zo arikerru (zo emphatic particle + aru “be” + keru subjectivizing suffix [attributive]), kara ni wa (kara “from” + ni “to” + wa topicalizing particle) “precisely because,” –nikeru kana (-nu perfective suffix + keri subjectivizing suffix [attributive form] + kana reflective exclamation), ni koso arikere (nari [copula split into ni ari], interrupted by koso [emphatic particle] + kere [conjunctive form of subjectivizing particle]), ma ni ma ni (ma “interval” + ni “to, in,” repeated) “while”

B. Predicates with Specific Words at their Core

1. Words built upon *akazu* “not get tired of”:
akazu “not get tired of” (*aku* “get tired of” + *a* [negative stem] + *zu* [negative suffix]), *akazu shite* “not getting tired of,” *akade* “without getting tired of,” *akanu* “not get tired of” (attributive form).

2. Words built upon *au* “meet”:
awade “without meeting,” *awamashi* “would meet,” *awamu* “meet” (tentative form), *au koto* “the act of meeting,” *ausaka no seki* “the border gate at Ausaka (the place name sounds like “slope where [we] meet”), *au yo* “I will meet” (emphatic), *au yoshi mo ga na* “if only there were a reason to meet.”

3. Words built upon . . . *ni izuru* “emerge in…:
iro ni ide “coming out in colors,” *iro ni wa ideji* “would not come out in colors,” *ho ni idete* “coming out in ears [of grain].”

4. Words built upon *omou* “think, long for”:
omowazu “without realizing it,” *omowanu toki* “when not thinking,” *omowamashi* “I would think,” *omowamu hito* “the person I would long for,” *omoioki* “keep in mind,” *omoikiya* “did I think?” *omoiki yu* “since I thought,” *omoikemu* “must have thought,” *omoikeru* “apparently thought” (attributive form), *omoihisome* “secretly longing for,” *omoinuru* “came to mind” (attributive), *omou kokoro* “longing heart,” *omou koro ka na* “O, the girl I long for,” *omou hito* “the person I long for,” *omoedomo* “I think, but…” *omooyu* “come to mind,” *mono wo omou* “think about things,” *hito wo omoi* “longing for a person,” *hito wo omou* “long for a person.”

5. Words built upon *kayou* “travel back and forth”:
kayoiji “the road that one travels on,” *kayoite* “traveling,” *kayou* “travel,” *kayoeru* “is traveling.”

6. Words built upon *kou* “be in love with”:

koishikari “was loved,” *koishikarikeru* “that which once was loved,” *koishikaru* “that which is loved,” *koishikaru beki* “ought to be loved,” *koishiki mono wo* “although [someone] is loved,” *koishi to* “when/if [someone] is loved,” *koitsutsu* “although in love,” *koi wa* “as for love,” *koi wa shi* “love is death,” *koimu to* “when about to fall in love,” *koi mo suru ka na* “I wonder if I should fall in love,” *koiwataramu* “shall I continue loving,” *koiwataru* “continue loving,” kouru “love” (attributive).

7. Words built upon *shiru* “to find out, know”:
shiranedo “I don’t know, but. . .” *shiramashi* “I would know,” *shirarezu* “is not known” (predicative), *shirarenu* “is not known” (attributive), *shirinuru* “find out,” *shiru hito* “a known person,” *shirubeku* “so as to know,” *shiruramu* “may know.”

8. Words built upon *miru* “see”:
mieshi “was visible,” *miezu* “is not visible,” *mienamu* “shall be visible,” *mienu* “become visible,” *miene* “become visible” (special predicative form), *mienedo* “is not visible, but…” *miewataru ka na* “I wonder if it might be visible from here,” *mimu hito* “the person I would like to see,” *mi mo senu* “doesn’t even see,” *miru made* “until [someone] sees,” *miruramu* “may see,” . . .*to miyuramu* “may look like…” …*to miru made* “until I see that…” …*to mo miezu* “doesn’t even look like…”

These are some of the examples that appear only in the works by the male poets. To use a literary work like the *Kokinwakashuu* as material for linguistic research, it should be treated seriously based on its special literary qualities, especially because these poems are clearly labeled with the name of the poet. This aspect allows the gender of the writer to be determined, which can be extremely useful for considering the language and gender relations of the classical period.

Previous studies have examined the advantages and disadvantages of using the *Kokinwakashuu* for these types of analyses and the history of this type of research [14]. However, as seen above, when comparing all character strings in the poems by male and female poets by using the n-gram set operator method, the nouns (as well as verbs with conjugational endings, adjectives, and negative and tentative suffixes) are extracted in a variety of compound word forms. Furthermore, some of these word forms assume certain cohesion in groups that have a single word as their core. With regard to defining words as units that assume both grammatical meaning and function, Nitta Yoshio (1997) states, “What actually appears in a sentence is not a word but a word form.”[15] In this case, he focuses his attention on groups of word forms that constitute sentences (such as the grouping of paradigmatic word forms *otoko wo* “man (direct object),” *otoko no* “man’s,” and *otoko ni koso* “indeed to a man”). The items in the previous table of “Predicates with Specific Words at their Core” can be referred to in other words as “groupings of paradigmatic word forms centered on specific words.” That is, comparison of all character strings with the n-gram set operator method provides an exhaustive sampling of occurring word forms from the corpus or text, which can be utilized in future research.

When we consider the additional axis of phase difference, the uses and semantic aspects of the extracted word forms become even easier to perceive. For some reason, examples of compounds with *shiro/shira-* “white” such as *shirayuki* “white snow” and others mentioned in the table above are concentrated in the *Kokinwakashuu* poems that were written by men. In addition, the same is true for words such as *akazu* “not grow tired of,” *au* “meet,” . . .*ni izu* “appear in,” *omou* “think, long for,” *kayou* “travel back and forth,” *kou* “be in love with,” *shiru* “know, find out,” and *miru* “see.” Words such as *shiru* and *miru* (which involve perception of the object), *kou* and *omou* (which express the subject’s thoughts), and *kayou*, which denotes motion toward an object, are concentrated among the male poets. Therefore, these words and forms can be categorized on the basis of gender differences.

When examining words in this manner, we can consider word forms based on *akazu* “not grow tired of” as an example. *Akazu* and its related forms stand out as words that tend to be found only in poems written by men. In addition, all of them are composed of the verb *aku* “grow tired of” as negative particles or suffixes. Naturally, we have always imagined that because *aku* and its noun form *aki* “boredom” existed during the Heian Period (the so-called heyday of classical Japanese), there would have been no difference between men’s and women’s use of these words. However, there are five examples of *aku* in poems written by anonymous authors, two of which are puns on *aku* “scum.” The nominal form *aki* “boredom” exists only as a pun on the word *aki* “autumn,” which gives a strong impression of being a special form that appears only as an aspect of traditional poetic rhetoric. If we expand our purview to the *Genji monogatari* (*The Tale of Genji*) and examine all occurrences of *aku* and its forms, what is especially notable is that *aku* generally appears in negative forms such as *akazu* “not grow tired of,” *akade* “not getting tired of,” *akanaku ni* “even though I do not grow tired of it,” and *akazarishi* “did not grow tired of.” Moreover, the majority of the subjects of *akazu* and other negatives are male, with few examples written by female subjects.

The forms of this verb are also used in a limited number of ways. Breaking down the 154 examples of the verb *aku* found in the *Genji Monogatari,* we find that 143 of them are in one of the negative compounded forms and only three are in other forms. In addition, the subject of the verb is male in 105 cases and female in only 19 cases, with the remaining 21 cases having phrases such as “people” or “the world.” Furthermore, not only are examples involving women rare but five of the examples include a mother speaking to her son, a grandmother speaking to a grandchild, or some other parent–child interaction, whereas six examples involve the dead. In other words, the words are used in interaction among people of different ranks, older and younger family members, the living and the dead, and others who are in a vertical hierarchical relationship in the social system. The fact that the higher-ranking person uses it is significant, and as a result, we can assume that the ratio of male to female users reflects society’s preference for males.

Thus, the question of how to view the actual situations, phase differences, and significance of the words that make up language has been made specific. There are many highly significant phenomena connected with the sampled compound words, but in the following, we will present the word form *haru no yamabe* “mountainous area in springtime” as another problem connected with compound words.

4. Semantic Compounds: *Haru no yamabe*

In syntactic terms, the phrase *haru no yamabe* "mountainous area in springtime," which has been identified only in poems written by men, is a noun phrase composed of noun plus case particle plus noun with additional information. It is usually not viewed as a single word; neither does it appear in Japanese dictionaries nor in major recent reference dictionaries of poetic language such as Katagiri [16] or Baba & Kubota [17]. According to the linguistic instincts of contemporary speakers of Japanese, this noun phrase is not worthy of any particular attention, but given that examples of its use are found only in the works of male poets in the *Kokinwakashuu*, it demonstrates that it ultimately functioned as a compound word.

We will now focus on examples of *yamabe* "mountainous area" from the *Man'yooshuu* (compiled ca. 759 AD):

Kasugano no yamabe no michi wo osorinaku kayoishi kimi ga mienu koro ka mo "My lord who came fearlessly over the mountain road of Kasugano is here no longer" (4-518) by Ishikawa no Iratsume (female).

Hisakata no ame no furu hi wo tada hitori yamabe ni oreba ibusekarikeri "Since I am alone on a rainy day in the mountains, I am depressed" (4-769) Ootomo no Yakamochi (male).

Nubatama no yoru wataru tsuki wo tomemu ni nishi no yamabe ni seki mo aranu ka mo "I try to stop the moon as it passes through the night, but I wish there were a travelers' checkpoint in the western mountains" (7-1077).

Sekigoshi ni inu yobikoshite togari suru kimi aoyama ni ha shigeki yamabe ni uma yasume kimi "Beyond the fence, calling his dog, my lord who is hunting with a falcon, let your horse rest in the green and leafy mountains of Aoyama" (7-1289).

There are many examples of *yamabe* in the *Man'yooshuu*, but there is not a single example of *haru no yamabe*. However, when exploring the examples of how *yamabe* is used, we find a place name, *Kasugano no* "of Kasugano," a direction, *nishi no* "western," and an adjective, *ha shigeki* "green and leafy." *Yamabe* thus functions as a noun that can be joined with any other word with no fixed collocations. There are also examples from female poets such as Ishikawa no Iratsume but we find no tendency for male poets to predominate. Yet in the *Kokinwakashuu* (compiled in ca. 905 AD), the formation of *haru no yamabe* "mountainous area in springtime" appears in addition to the following types of examples found in the *Man'yooshuu*:

Iza kyou ha haru no yamabe ni majirinamu, kurenaba nage no hana no kage ka ha "Come today and let us get lost in the springtime mountains, and if the sun sets, we might shelter ourselves under the blossoms" (Spring 2-95) Sosei (male).

Kasumi tatsu haru no yamabe wa tookeredo fukikuru kaze wa hana no ka zo suru “The springtime mountains where the mist rises are far away but the wind blowing down from them bears the scent of blossom” (Spring 2-103) Motokata (male).

Azusa yumi haru no yamabe wo koekureba, michi mo sariaezu hana zo chirinikeru “Crossing the mountains in springtime when the days grow long as the catalpa bow I am lured off my path by the drifting blossoms” (Spring 2-115) Tsurayuki (male).

Yadori shite haru no yamabe ni netaru yo wa yume no uchi ni mo hana zo chirinikeru “As I lodge, sleeping in the springtime mountains, the blossoms scattered, even in my dreams” (Spring 2, 117) Tsurayuki.

Omou dochi haru no yamabe ni uchimurete soko to mo iwanu tabine shiteshika (Spring 2-126) “Oh that we might gather in the springtime mountains and sleep outside telling no one where we are” (Spring 2-126) Sosei.

All of the poets cited here are male, and each of the examples are seen in the volume titled “Spring 2.” Furthermore, these examples share common phrases and meanings: the men “go to” or “get lost” or “sleep in” the “springtime mountains” colored by “blossoms.” These phrases suggest images of promising to spend one night with a woman, and the phrase *haru no yamabe* functions as a metaphoric symbol of the woman who is the object of that promise. This phrase is not just a so-called transient catchphrase found only in the *Kokinwakashuu*, but it continued to be used in the same symbolic sense in subsequent years.

Next, we compare all character strings in the poems written by the female poets:

Wa ga yadorishi haru no yamabe no tsuma nareba hoka no hana to mo omooenu ka na.” When you are my spouse in the springtime mountains where we spent the night, I don’t even think about the flowers” *Nakatsukasashuu*, no. 37 (The *Nakatsukasashuu* is a tenth-century anthology).

Yogoto tada tsukuru omoi ni moewataru wa ga mi zo haru no yamabe naramashi “If only I, who burn with passionate thoughts every night, could become a springtime mountain” Daini no Sanmi Shuu, no. 50. (Daini no Sanmi was a female poet of the eleventh century.).

In poems such as these, we find scattered examples in which “where we spent the night” is part of the “springtime mountains” and the term “I” is also a “springtime mountain.” On the other hand, the fact that no examples by male poets equate “springtime mountain” with themselves is a reflection of this metaphorical meaning and gender difference in the use of this phrase. The *Kokinwakashuu* was compiled in 905 AD and the *Nakatsukasashuu* was written in the middle of the 10th century. In addition, Daini no Sanmi was a poet who lived in the first half of the 11th century. Therefore, for more than a century after the compiling of the *Kokinwakashuu*, *haru no yamabe* was used with a specific meaning as a cohesive phrase. Given that this semantically strong collocation lasted for more than a century, it may be appropriate to acknowledge it as a compound word. In this case, we would like to treat all word formations like these as compounds newly created as filters for social and cultural settings and phenomena or compound words in the semantic sense.

Haru no yamabe is from the language belonging to the literary domain of *waka* poetry. However, the question is whether it is possible to take certain groups of words from the past and contemporary linguistic phenomena and find their commonalities. As Yamamoto Kiyotaka [18] points out, *aoi yama* "green mountain" is a phrase that symbolizes happiness and *akai hane* "red feather," as in the Red Feather [Charity] Campaign, have the same types of characteristics. Their forms and accentuation patterns make it difficult to refer to them as compound words, but to the extent that they are used in specific meanings, they must inevitably be considered as single compound words. In previous research, there has been no option other than to discover these expressions arbitrarily, and because it is extremely difficult to sample them exhaustively, questions of how to classify them semantically or what position they hold among compound words as a whole have shown no definitive answers. Yet, we can assume that there are a considerable number of these particular compounds in both classical and modern Japanese. If we can identify word forms from comparisons of all character strings in corpuses with different phases (not only for gender differences but also for differences in speech across age groups, differences between early modern and more recent literature, and other types of phase differences) and apply analytical techniques to them, a variety of semantic compounds will be discovered. As a result, we will be able to state that certain word formations function as a semantic compound only in certain phases.

5. Conclusion

In the preceding pages, we have proposed rules for Japanese linguistic structures based on regular grammar (type 3 grammar) that takes an approach different from phrase structure grammar (type 2 grammar), which is the current mainstream theory that examines linguistic structures in terms of syntax. We have focused on the emergence of these rules and showed that utilizing n-gram statistics is an appropriate technique for the exhaustive sampling of texts and corpuses. Moreover, we believe that by encompassing all of the actual word forms, we have succeeded in expanding our knowledge about words and compounds in terms of grammatical theory, semantics, and phase theory.

Notes:

[1] Kokuritsu Kokugo Kenkyuujo, Kokuritsu kokugo kenkyuujo houkoku 21, *Gendai zasshi kyuujuushu no yougo, youji* (Part 1: "Souki oyobi goihyou"), 1962.
____Kokuritsu kokugo kenkyuujo houkoku 25, *Gendai zasshi kyuujuushu no yougo, youji*, (Part 3) 1964.

[2] Sakakura, Atsuyoshi, *Go kosei no kenkyuu,* Kadokawa Shoten, 1966.

[3] Kubokin, Yasuo *Go keisei to on'in kouzou*, Kuroshio Shuppan, 1995.

[4] Okitsu, Keiichirou, "Fukugo meishi no seisei bunpou," *Kokugogaku* 101, Kokugogakkai 1975. Nitta, Yoshio, *Goironteki tougoron*, Meiji Shoin, 1980. Kageyama, Tarou, *Bunpou to gokeisei*, Hitsuji Shobou, 1993.

[5] Miyaji Hiroshi, *Keigo, jouyouku hyougenron: gendaigo no bunpou to hyougen no kenkyuu* (2), Meiji Shoin, 1999. Yamamoto Kiyotaka, "Tanjungo, fukugougo, hasseigo," *Nihongogaku,* 14-5, Meiji Shoin, 1995.

[6] Namiki, Takayasu, "Fukugougo no Nichi-Ei taishou: fukugou meishi, fukugou keiyoushi," *Nihongogaku* 7-5, Meiji Shoin, 1988. Kageyama,1993. Kageyama Tarou and Yumoto Youko, *Go keisei to gainen kouzou*, part of the series *Nichi-Eigo hikaku sensho*, Kenkyuusha Shuppan, 1997.

[7] Miyajima, Tatsuo. "Sousakuin e no chuumon," *Kokugogaku* 76, Kokugo Gakkai 1971. Ishii, Masahiko, "Jisho ni noru fukugougo, noranai fukugougo," *Nihongogaku* 7-5, Meiji Shoin, 1988.

[8] Yamamoto, Kiyotaka "Fukugougo to bun no kyoukai," *Nihongogaku* 15-9, Meiji Shoin, 1996.

[9] Okitsu, 1975.

[10] Mizutani Shizuo, *Kokugogaku itsutsu no hakken saihakken*, Toukyou Joshi Daigaku Gakkai, 1974.

[11] Nagao, Makoto, Mori Shinsuke, "Dai kibo Nihongo tekisuto no n-guramu toukei no tsukurikata to goku no jidou chuushutsu," *Shizen gengo shori* 96-1, 1993. For the purposes of this study, both Professor Nagao Makoto, Chancellor of Kyoto University, and Mori Shinsuke of IBM Japan's Tokyo Basic Research Institute, graciously allowed me to use software that they had developed for high-speed processing of n-gram statistics. I am taking this opportunity to note my profound gratitude.

[12] Shannon, Claude E. and Warren Weaver. *The Mathematical Theory of Communication*, the University of Illinois Press, 1949.

[13] Kondou, Miyuki, "Heian jidai waka shiryou ni okeru tokushu goi chuushutsu ni tsuite no keirouteki kenkyuu to riyou tsuuru no koukai: Kokinwakashuu no kago to hyougen no jendaasei ni tsuite," *Kagaku kenkyuuhi tokutei ryouiki kenkyuu, jinbun kagaku to konpyuuta, kenkyuu seika houkokusho, konpyuuta shien ni yoru jinbun kagaku kenkyuu no suishin: 1999*, 1999.
______ "N-guramu toukei shori wo mochiita mojiretsu bunseki ni yoru Nihon koten bungaku no kenkyuu: *Kokinwakashuu* no 'kotoba' no gata to seisa," *Jinbun Kenkyuu* 29, Chiba University, 2000. Kondou, Yasuhiro and Kondou Miyuki, "Heian jidai kotengo koten bungaku kenkyuu no tame no n-gram wo mochiita kaiseki shuhou," *Gengo shori*

gakkai dai 7 kai nenji taikai happyouron bunshuu, 2001. These are some of the studies in which I found the details of the framework for set operation methods.

[14] Kondou, Miyuki, 1999 and 2001, as well as “Kokinwakashuu no ‘kotoba’ no gat: gengo hyoushou to jendaa,” presented at 2001 annual conference “The Genesis of Gender,” at the Kokubungaku Kenkyuu Shiryoukan. Available at http://www.nijl.ac.jp/events/openlecture/01kouenkai.htm. Published by Rinsen Shoten, 2002.

[15] Nitta Yoshio, *Nihongo bunpou kenkyuu josetu: Nihongo no kijutsu bunpou wo mezashitte,”* Kuroshio Shuppan, 1997.

[16] Katagiri, Youichi, *Uta makurakotoba jiten,* (expanded version), Kasama Shoin, 1999.

[17] Baba Akiko, and Kubota Atsushi, *Uta kotoba, uta makura jiten,* Kadokawa Shoten, 1999.

[18] Yamamoto, Kiyotaka, 1955.