Japanese Literary Research Using Corpuses: The N-gram Method and “Linguistic Resources”

Miyuki Kondo

1. Introduction

In 2000, I developed a research method called “character string set operation method with N-gram” (1), and I conducted a variety of research using this method. Also, in recent years, the application of this method has proven that it is possible to perceive the phenomenon of specific language expressions being accumulated in a type of literary work, its aggregate, or language group, and to enjoy and further reproduce and spread expressions. These types of linguistic data, containing specific linguistic expressions, are called “linguistic resources” in sociolinguistics (2); however, as noted below, a variety of literary phenomena can be explained using the concept of linguistic resources. Particularly, through using a corpus and research methods such as the N-gram method, the actual condition of linguistic resources can be explored, and through this, a new and unconventional research discipline can be pioneered.

Traditionally, in the Japanese literature and language linguistics research field, research methods employing the N-gram method had not come into common practice, and to a certain extent, there was a lack of understanding as to what kind of research it was. Accordingly, in this paper, I would first like to reintroduce corpus research with the N-gram method, and through this new concept of linguistic resources coming to light, I would like to demonstrate the completely new knowledge acquired from Japanese literature and Japanese language linguists.

Author 2018/3/7 12:06

コメント **[1]:** Remark: Note that the current expression “Japanese literature and language linguistics research field” is as per the original text. However, it is unclear. Please check if “Japanese literature and language research field” or “Japanese literature and linguistics research field” should be used for clarity and accuracy. Please check and clarify further if required. Please check the phrase at the second instance as well. We have left a yellow highlight for your reference at the instance of second usage of the expression.

2. Corpuses and N-gram

“N-gram” refers to the act of collecting a combination of units, such as phonemes, characters, or words, from a subject’s linguistic data, in one-unit groupings (1-gram), two-unit groupings (2-gram), three-unit groupings (3-gram), and on to N-unit groupings, or it may refer to the collection itself. They are generally used as indexes, and their data characteristics can be measured. Particularly, in terms of their characteristics, character N-grams can completely encompass all the character string patterns in the data, so an index can be made that is reliably more comprehensive than one with words extracted from morphological analysis. Accordingly, it is widely used in search engines and situations where comprehensive searches are needed.

The “character string set operation method with N-gram” referenced in the introduction is a method I developed for classical literary work analysis, in which the entirety of a classical literary work is converted to Hiragana; N-grams with character units are gathered and then mutually compared with set operations in several different texts. Through this, for example, if collected into a 10-gram, since all character strings in a range from one-character to ten-character can be compared, the research referenced in the beginning and the poems of the male and female poets of *Kokinshu* (abbreviated name of “Collection of Poems of Ancient and Modern Times”) can be mutually compared, and by seeking out the different sets, the exclusively male (or female) words and expressions can be sought (3). There also exists research that separately evaluated character strings that are common to *Kokinshu* and *Genji Monogatari* (*The Tale of Genji*, a common set), namely items corresponding to words, collocations, and *hikiuta* (analogies) (4). Hiragana-unit character string N-gram could linguistically be considered phonologically mora N-grams, so it is an extremely efficient method for analysis of the Japanese language. It is also highly effective for compound-word research (5). There is also a method of sampling an N-gram of phonemes rendered in romaji, analyzing the phonemes, and using a word N-gram, but I will omit that here.

Generally, a corpus morphological analysis (breaking down words into lexemes such as parts of speech, phonetic reading, and restoration of conjugation to original form) is done, and the results of this analysis are embedded in the corpus as annotations, and a KWIC concordance is made using that information. However, with a character string N-gram, this type of annotation is completely unnecessary, and if the phonetic reading of the kanji characters has been completed, they can be converted into Hiragana, and an N-gram can be made.

When this technique was originally conceived in the 1990s, it was not even commonplace to quickly acquire an N-gram with a computer, and research began under instruction from both Makoto Nagao and Shinsuke Mori who developed a high-speed algorithm for this purpose (6). Yet, considering the present, N-grams have become ubiquitous in the world of computer science and one cannot help feeling the progress over the years.

3. The Pursuit of “Linguistic Resources” Through N-gram

Incidentally, the first research I conducted through the character N-gram mutual operation described in the introduction was vocabulary and gender expression differences seen in *Kokin Wakashu* (formal name of “Collection of Poems of Ancient and Modern Times”) (7). There were many specific expressions used only by male poets, such as the verbs *akazu*, *miru*, *shiru*, and *kofu*, and phrases used to express scenery, such as *wo minaheshi* and *haru no yamahe*; therefore, I stated that this was closely linked to restrictions as an element of the Nijūichidaishū (Collections of the Twenty-One Eras) of the *Kokin Wakashu*. Further, this type of inclination toward gender-specific terminology is evident not only in *Kokin Wakashu* but also, naturally, in later Nijūichidaishū, and I discovered that this division is distinctly visible in the waka contained in the stories of *Genji Monogatari*. In short, this indicates that *Kokin Wakashu* formed the source for the gender standard, and future-generation literary works continued to use and reproduce it to generate works.

Author 2018/3/7 11:40

コメント **[2]:** Remark: Note that the highlighted text lacked clarity when translated into English. To improve clarity and readability, we have modified certain content in the highlighted text. Please check and clarify further if required.

These types of sources for linguistic information linked to a specific identity are called “linguistic resources” in American sociolinguistics. Momoko Nakamura introduced it in her modern Japanese gender linguistics research, translating and presenting it as *gengo shigen* (linguistic materials) (8), but *gengo shigen* is used in natural language processing to mean language resources and data in a very general sense, so the editor intentionally uses the translation of *gengo risoosu* (linguistic resources) (9). I would like to follow that translation in this report.

From a “linguistic resource” standpoint, *Kokin Wakashu* continued beyond the Heian Period into future ages to be used as a standard for verse form. As signified in the word *kokin denju* (the teaching method of how to interpret the *Kokin Wakashu* passed from generation to generation), the words and phrases included within were even said to be mystic teachings and were passed down. Accordingly, in the Heian Period, of course, waka were made with importance given to resources such as *Kokin Wakashu*. This means that knowledge can be gained about a variety of linguistic phenomena from the aspect of how these resources were "handed down" and "reproduced," such as in the case of gender; however, at the same time, the aspects of "transformation" and "deviation" cannot be overlooked.

One good example of "transformation/deviation" is the *Sanbyakurokuju-shu Uta* ("Three Hundred-Sixty Poems," also called *Maigetsushu*) by Sone no Yoshitada. As I have already indicated, and will describe in the next section, there are many areas where Sone no Yoshitada appears to diverge from the *Kokin Wakashu* standard, and this can be considered one aspect of this poem collection's literary significance (10). In the next section, I would like to view those characteristics from literary and linguistic perspective.

4. The Aspect World of *Sanbyakurokuju-shu-Uta*

Sanbyakurokuju-shu-Uta "Maigetsushu" is a *teisukashu* (collection of a predetermined number of poems) composed by Sone no Yoshitada during the reigns of Emperors Murakami and Enyu, half a century after the formation of the *Kokinshu*. In this period, beginning with Yoshitada's *Hyakushu-Uta* ("One Hundred Poems"), Minamoto no Shitagou, Minamoto no Shigeyuki, and others had early *teisuka* that had around 100 poems, called *kagun* (poem groups), but the *Sanbyakurokuju-shu-Uta* can be considered the most important of these works. In the poems of Sone no Yoshitada, original materials that draw on *Manyoshu* (poem anthology, "Collection of Ten Thousand Leaves") and Chinese poetry as their source are used, and that these terms are unique is a long-known fact, but I used an additional analysis method through N-gram to conduct a mutual comparison of all character strings from 1- to 5-gram among a group of *teisuka* from the same period—*Yoshitada Hyakushu* ("One Hundred Poems of Yoshitada"), *Shitagou Hyakushu* ("One Hundred Poems of Shigato"), and *Shigeyuki Hyakushu* ("One Hundred Poems of Shigeyuki")—and extracted the character strings original to *Sanbyakurokuju-shu-Uta*. Examining this, I found new nouns and proper nouns like *Asajihe* and *Atagoyama* that did not appear in *Kokinshu*, but the major characteristic was that within language that should be called, in a broad sense,

Author 2018/3/7 11:46
コメント [3]: Remark: Our translation at this instance is as per the source text. Because it is unclear in the given context, we recommend that you clarify this through our Q&A service.

expression of time (verbs, adjectives, particles, and auxiliary verbs), many original expressions were included. I show below a list of these using my own categorization.

I. Word Forms Consisting Primarily of Independent Words

- Words expressing a point in time

yube, no yugure, hiyori ni, hirune, hiruma, uzuki, nagatsuki, yayohi, tokizo shiru

- Words expressing a starting point in time

ashita yori, o kyo mireba, aki no hatsukaze, ima sara ni, imazo, sono kami, taeshi yori, hatsuaki, hatsune, fuku kara ni, and *fuku nahe ni*

(*Nishi*) + time (not in any other *Hyakushu*) + (*Yori*):

nishi ashita yori, *nishi yube yori*, *nishi hi yori*, and *nishi sono hi yori*

- Words expressing an ending point in time

akihatete, haru no kure, natsu no kure, natsu no kururu, ohari

- Nouns, adverbs, and compound words expressing the passing of time

hizonaki, mataki, ni tsuketezo, ni tsukete, ni tsuketemo, toshitsuki, yogoto ni, asanayufuna ni, itsu shika na, itsu shika mo, ihanu hi, sugigate, suru hodo (long-verse poetry), *nagaki yoru, natsugoto ni, madashiki, yonayona*

- Verbs expressing the passing of time

isogi, isogu, isoge, itomanami, usuragi, toshifureba, toshi o tsumi, narinureba, ni sugushi, nite akasu, o kazohe (toshi no nissu wo kazohe, ikuso tsukihi o kazohe), akegataki, ikuyohenu, gate ni suru, kiesenu, sanaheoi, suginuran, tsukurioki, tsumoruran, tsumoru o, toshitsumori, nagabiku, nagamuru, nurubami (long-verse poetry), *moenokoru, yasurafu, o rikurashi*

- Adjectives that express the passing of time

nagaki (in long-verse poetry), *nagashi, nagaku, matoho, himanaki, himamonaku*

II. Word Forms Consisting Primarily of Ancillary Words

- Particles

Tsutsu: *shiguretsutsu, tsumoritsutsu (chiritsumoritsutsu, yukitsumoritsutsu), furitsutsu, okitsutsu, omokakushitsutsu, oritsutsu, kuyuritsutsu, tatohetsutsu* (long-verse poetry), *toshi o tsumitsutsu, nagekitsutsu, natsuketsutsu (mitoseohinokoma), natsuketsutsu* (long-verse poetry), *midaretsutsu, yukitsutsu, yokitsutsu, yosohetsutsu, waketsutsu, o rikurashitsutsu*

Te: *sashite, fumiwakete, tokete, akihatete, usurete, uchimurete, oritachite, namitachite, machikanete, o sashite, akeokite, atsurete, usuragite, uchitokete, uzumorete, kakesagete, koritsumite, sasoharete, shinobite, sekitomete, te ni torete, tojirarete, nazukete, nokoshi o kite, hijite, himouchitokete, matohonite, makase o kite, mirete, minarete, yanasashite, haruyamakete, yo o hete, o kite*

- Auxiliary Verbs

Nikeri (includes *nikeri, nikeru, nikere*) : *irozukinikeri, usuraginikeri, utagahinikeri, katabukinikeri, karehatenikeri, kienikeri, shigeriahinikeri, tachinikeri, harukinikeri, hikobaenikeri, miminarenikeri*
Tsu: *tehedatsuru*
Nu: *sashoharenu, samenu, chikazukinu*
Tari: *tsukuritaru, mureitaru*
Ri: *kakareri, sahoseri, saraseru, nurumeru, nokoseru, musuberi*
Nishi: *karenishi, kurenishi*

III. Compound-word forms

● Compound verb suffixes

- go (*yuku*): going to happen (*okiteyuku*), going to gradually get cold (*saeyuku*), going to point (*sashiteyuku*), going to get (*toriniyuku*), go in the end (*yohariyuku*)
- come (*kuru*): when summer comes (*natsukureba*), in the coming spring (*harukini*), come into view (*miekuru*), if someone comes to share (*wakekureba*)

Compound-word forms with going/coming (*yuku/kuru*): person who is coming (*kuruhito*), if it comes (*kitaraba*), the fall that comes and goes (*kureyukuaki*), the road one goes (*yukumichi*), gradually progressing (*nariyukusama*)

- cross over/cover (*wataru*): cross over (*wataru*), covered in green (*awomiwataru*), cross over from sleep (*okiwatari*), covered in mist (*kasumiwatarishi*), has listened over a long period of time but (*kikiwataredo*), when covered in clouds (*kumoriwatareba*), when covered in desire (*kogarewatareto*)
- while (*hodoni*): while thinking profoundly (*omohishihodoni*), while airing out (*sahosuhodoniso*), while in the grass (*susukihodoni*), while doing (*suruhodoni*), while they were (*toseshihodoni*), while being prolonged (*nakabikuhodoni*), while gazing (*nagameshihodoni*), while looking (*mishihodoni*), while not seeing (*minuhodoni*)
- finishing (*hatsu*): fall ending (*akihatete*), finished living (*karashihatete*)
- to be (*bamu*): being painted (*nurubami*) (from a long Japanese poem)

Author 2018/3/7 12:06

コメント **[4]:** Remark: Our translation at this instance is as per the source text. Because it is unclear in the given context, we recommend that you clarify this through our Q&A service.

Looking at this data, it becomes apparent that there are an overwhelming number of word forms that express aspects of time passing or continuing. When the data is limited to independent words, there are many variant word forms whose attributes express the passage of time rather than a simple point in time. Some of these variants are “as the day ends” (*hizonaki*), “another time” (*mataki*), “continuing through” (*nitsuketezo*), “continuing” (*nitsukete*), and “even after continuing” (*nitsuketemo*). It also becomes obvious that many forms of ancillary attached words expressing continuation or completion appear as unique expressions in the *Collection of 360 Poems*. Conversely, *ki* and *keri*, which indicate the past and reminiscence, are not extracted as unique word forms; they are extracted only in the forms *nikeri* and *nikeru*, when they are connected to the perfect negative form of *nu*.

At this point, one specific method of organizing time in poems that use time expressions will be considered.

○The thorny plants look at the river as they age, while we also accumulate the years in bitterness. (107/mid-April)

○ Frozen dew arises in the night; while looking (*mishihodoni*) at the winter night's moon, the tears on my sleeves freeze (*kohorinu*). (322/mid-November)

As these examples suggest, during the repetition or continuation of the thoughts or actions of their subjects, these poems use functions that can be regarded as expressions indicating the passage or accumulation of time. In the first poem, the inverted word order of “while years are accumulated” and “they age” serves this function, and in the latter poem the passage of time and its completion are found in the single lines “while looking (*mishihodoni*)” and “freeze (*kohorinu*).” Put simply, these expressions represent time as an aspectual expression.

Author 2018/3/7 12:06
コメント [5]: Remark: Our translation at this instance is as per the source text. However, it is unclear in the given context. Please check if “one-word phrases” would be better in place of “single lines.”

The verb “rushes” (*isogu*) and the noun “rush” (*isogi*) should be mentioned as highly characteristic of these sorts of expressions. Of all the sources that contain early poems with set numbers of words, these verbs appear only in the *Collection of 360 Poems* and there are no examples in the *Kokinshuu*. Furthermore, they appear in the following seven poems in the *Collection of 360 Poems*.

○Leave it to a person who has seedlings in them, I will rush to see the flowers (58/end of February)
○ If your heart is colored by seeing the cherry blossoms, spring will rush to set out in full (72/beginning of March)
○ The protector of the Mita shop will rush to quickly plant seeds, as it is May today, or they will wilt (125/beginning of May)
○The night's dew will be left behind somewhere, and the leaves of the rice plants will make people rush (199/mid-July)
○ When struck by the wind when clothes have not been mended, the deep fall nights rush to the cold (240/end of August)
○ The yearning covers the open sea, but when it is winter, the sailors will rush through the cold sea routes (326/mid-November)
○ Will I be rushed to be discouraged and have no free time (*Tamesukehon*)? The soul winter was truly spoken of (367/end of December)

“Rushing” limits the amount of time for which an action is being done, and it is also an aspectual expression in the broader sense. In this way, the world depicted in the *Collection of 360 Poems* is overwhelmingly created using language that expresses the passage of time: the aspects of time that are intrinsic to the subject of the poem. An overall comparison with the *Kokinshuu* is planned for the next stage of research, but these examples using the term "rush" provide a glimpse of a deviation from the ancient order.

As has been described previously, the unique language in the *Collection of 360 Poems* contains many aspectual expressions. It was also noticed that the previous list contains words that do not simply indicate the passage or continuation of time. Examples include the compound verbs "come" and "go" in the phrases "go to get" and "if one comes to share." These words can be used to indicate time—as in "go in the end" and "when summer comes"—but they were originally expressions indicating space. In addition, the suffix compound verb "crossover" can be used for time, as in "covered in desire," and for space, as in "covered over in mist."

Because these compound verbs are only connected to specific verbs, they can be thought of as lexical compound verbs. Certain suffix items on the list, such as "finishing" and "to be," are indications of time, whereas other items, such as "left behind," indicate space. Both types of words create a world with a lexicon characteristic of the *Collection of 360 Poems*, even among the various types of poems that use limited numbers of words.

A recent study by Taro Kageyama proposed a new two-category system for classifying lexical compound verbs into subject-related compound verbs, where the items before and after the verbs are related (such as "to push off," *tsukiotosu*) and aspect compound verbs (such as "hurrying to die," *shiniisogu*) (12). In addition, the (lexical) aspects of the latter can be categorized into temporal Aktionsart (aspects) and spatial Aktionsart (aspects). In line with these classifications, the compound verbs in the expressions characteristic of the *Collection of 360 Poems* include spatial descriptions and encompass aspect compound verbs.

Conversely, the fact that literary works such as the *Collection of 360 Poems* contain this particular sort of group lexicon corroborates the validity of this classification of aspect compound verbs.

5. The establishment of "linguistic resources"

This paper introduces the present author's theory, drawing on the fact that the *Collection of 360 Poems*, with its deviations from traditional "linguistic resources" such as the *Kokinshuu*, contains characteristic items that seem to belong to a single linguistic category (that of aspect compound verbs). This element of deviation makes it possible to gain insight into the nature of the *Kokinwakashu* as a "linguistic resource."

The completion of the *Kokinshu* in the Nijūichidaishū could be interpreted as signifying the "completion" of a cultural linguistic resource. Thus, linguistic information that has a variety of significance attached to it was completed at that point. The significance attached to this linguistic information includes the forms of men's and women's language (masculinity and femininity), its combination with the natural scenery (a nightingale on a plum tree), perceptions of space and time, and the way in which specific events and phenomena are tied to specific emotions (the sadness connected to an autumn moon and the smell of a tachibana connected to thoughts of parting with someone long ago). Some of these motifs, such as the connection of the autumn moon with sadness, have been extended to modern Japanese sensibilities, whereas others, such as the connection of the smell of a tachibana connected to thoughts of parting with someone long ago, ended with the noble society of that time.

Traditional gender theory calls linguistic resources "sources that provide linguistic information connected to the specific identity of being male or female," but this paper hopes to avoid limiting this definition to sex and instead extend its meaning to cover a broader area. Thus, linguistic resources are redefined here as "sources that provide linguistic information with specific significance attached to them," such as how natural scenery is perceived and how language users handle space, time, emotions, and perceptions. This seems to be a new concept in the study of literature and linguistics, made feasible for the first time by the possibilities afforded by corpus research.

Notes

1. KONDŌ, Miyuki, "An Analysis of Japanese Classical Literature Using Character-based

N-gram model-Differences seen in the word forms and gender in the *Kokinwakashu*" *Chiba University Journal of Humanities*, Issue 29, 2000. In this study, the present author conducted set calculations on two literary sources; in a subsequent study, Kosei Ishi'i proposed carrying out similar calculations on multiple literary sources and called the NGSM (N-gram Based System for Multiple Documents) technique. See also Ishi'i, Kosei, "N-Gram's feasibility: Comparing Different Texts in Buddhist Literature: Judging the Translators and Authors" *Journal of Japan Association for East Asian Text Processing*, Issue 2, 2001. Software for NGSM ngmerge has been developed by Yasuhiro Kondō (see http://japanese.g., r.jp/app-def/S-100/main/).

2 See Eckert, Penelope and Rickford, John R., *Style and Sociolinguistic Variation*, Cambridge University Press, 2001.

3. See note 1.

4. KONDŌ, Yasuhiro, "Digital Texts as 'Cultural Resources': Common Problems for the Japanese Language and Japanese Literature" *Japanese Language and Japanese Literature*, vol. 77, 11th Edition, 2000.

5. KONDŌ, Miyuki, "Extracting Word Forms Using N-gram Statistics and Compound Words, Based on an Analysis of Heian-era Japanese" *Japanese Linguistics*, vol. 20, issue 8, 2001.

6. NAGAO, Makoto and MORI Shinsuke, "How to Create N-gram Statistics for Large-scale Japanese Texts and the Automatic Extraction of Words" *Research Reports of the Information Processing Society of Japan*, vol. 93, 1993. The software that was used in this paper to extract N-grams was provided by both authors.

7. See note 1.

8. NAKAMURA, Momoko, *Gender and Japanese: Men and Women Created by Language*, NHK Publishing, 2007.

9. KONDŌ, Miyuki, "The Culture of the *Tale of Genji*: The *Tale of Genji* and Gender—The Men and Women Created from the Language of Poems" [speech] *Jissen Women's Educational Institute Bungei Material Laboratory Annual Report*, vol. 28, 2009.

10. KONDŌ, Miyuki, "The Construction of Anti-Kokin 'Behavior': Essay on Sone no Yoshitada's *Collection of 360 Poems*" Iwanami Shoten, *Literature*, vol. 6, issue 4, 2005.

[11] See the source in note 10 for a more detailed explanation of this table.

[12] KAGEYAMA, Taro, “A New System for Lexical Compound Verbs: The Theoretical and Practical Implications” in *Elucidating the Cutting-edge Mysteries of Research into Compound Verbs*, Hitsuji Shobo, 2013.

The paper is one of the outputs of the research in development of linguistic resources during the Heian Period funded by Grant-in-Aid for Scientific Research (B).

(KONDŌ, Miyuki, Professor at Jissen Women’s Educational Institute)