UDK 811.521’367.624
Izvorni znanstveni rad
Rukopis primljen 19. VII. 2024.
Prihvaćen za tisak 16. XII. 2024.
doi.org/10.31724/rihjj.50.2.9

**Irena Srdanović**
Juraj Dobrila University of Pula, Croatia
https://orcid.org/0000-0003-1281-176X
*isrdanovic@unipu.hr*

# DISTRIBUTION OF SUPPOSITIONAL ADVERBS IN JAPANESE LANGUAGE CORPORA: USING SKEWER-SEARCH SYSTEM KOTONOHA

This paper aims to analyze and describe the distribution of suppositional, more broadly, modal adverbs in Japanese language corpora using state-of-the art resources for empirical linguistic analysis. Specific adverbs with suppositional modality meanings, such as *tabun*, *osoraku*, *kanarazu*, *kanarazushimo*, *kitto* etc. are compared in their usage across multiple corpora using the concordance system *“KOTONOHA”* developed at the National Institute for Japanese Language and Linguistics. This analysis aims to indicate which adverbs are commonly used across a variety of contexts and which are more specialized. The study compares usage across eleven different Japanese language corpora, including written and spoken corpora, learners’ corpora, diachronic corpora, revealing differences in genres, registers, native speakers’ or learners’ usage, as well as changes over time. Two widely used adverbs, *tabun* and *osoraku* ‘probably’, belonging to the expectation modality type, are used as a model to present retrieved data and graphs with detailed interpretation. The research enhances understanding of adverbs’ linguistic features, contextual usage, and evolution, with implications for Japanese language education, while proposing a methodology for future studies using multiple corpora and suggesting advancements in language resources to stimulate further research.

# 1. Introduction

Adverbs represent a minor word class in the Japanese language compared to the much more represented nouns and verbs (Joyce et al. 2012), however, their communicative significance is essential, particularly as modifiers of other word classes, allowing words to convey the intended meaning more precisely. As such, they appear to be represented in various corpora and types of texts, often indicating the nature of the text itself and specific genres within the corpora, as was already pointed out in previous research on suppositional adverbs (Srdanović et al. 2008a,b, 2009a,b; Hodošček et al. 2009).

In the last two to three decades, Japanese language corpora and related resources have developed enormously fast, providing language researchers with highly relevant data for detailed linguistic analysis. Specifically, the National Institute for Japanese Language and Linguistics (NINJAL) has made significant efforts in developing various types of language corpora and resources, often collaborating with other institutions. The effort has resulted in tens of user-friendly and accessible data and tools. One such tool is the concordance system for searching Japanese language corpora, known as "skewer-search system KOTONOHA" (jap. *matomete kensaku "KOTONOHA"*) (Oka et al. 2020). It is a state-of-the-art concordance tool designed for searching and comparing multiple corpora by categories, such as mode of communication (written or spoken corpora) and various historical periods (from the Nara to the Heisei period).

The aim of this paper is to explore the distribution and usage of specific suppositional adverbs using the novel state-of-the-art concordance tool for searching various types of Japanese language corpora. The research questions to be addressed are: What is the distribution of specific suppositional adverbs and what are the differences in usage among various a) written and spoken corpora, b) genres and registers, c) native speakers' and learners' corpora, d) historical periods in diachronic corpora?

The second section discusses suppositional adverbs as a category within Japanese adverbs, referencing previous studies. The third section explains the data and methods employed in the analysis. The fourth section addresses the above-posed research questions and presents the analysis results, including the distribution of twenty-two suppositional adverbs across eleven Japanese language corpora, with a specific focus on the widely used adverbs *tabun* and *osoraku*. The final section summarizes the findings and suggests directions for further research.

## 2. Japanese modal adverbs and previous research

Various attempts have been made to classify Japanese adverbs into categories based on their diverse functions within sentences (see for example, Yamada 1936, Masuoka and Takubo 1989, Ishiguro 2023). The classification by Yamada (1936) is one of the most widely known dividing Japanese adverbs into: i) adverbs of degree (jap. *teido fukushi* 程度副詞) that specify the degree of an action or state, such as *sukoshi* 'a bit', ii) adverbs of manner (jap. *jōtai fukushi* 情態副詞) that specify the manner of an action, such as *yukkuri* 'slowly', and iii) statement or modal adverbs (jap. *chinjutsu fukushi* 陳述副詞), such as *tabun* 'probably', that express the speaker's subjective judgment or proposition and often correlate with the clause-final modality.

Kudô (2000) focuses his research on suppositional adverbs (jap. *suiryō fukushi* 推量副詞) as a subgroup of modal adverbs. His research used a smaller amount of data consisting of newspapers and modern literature, but clearly indicated tendencies in how modal adverbs exhibit a strong agreement-like behavior with clause-final or utterance-final modality forms, such as *hazu da* '(It) should be', *darō* '(I) think/suppose', *rashii* '(It) looks/seems', *kamoshirenai* '(It) may'. This observation aligns with Minami's (1974) description of the layered structure of Japanese sentences, also known as the nesting structure *ireko kōzō* (入れ子構造), reminiscent of Japanese or Russian nesting dolls (*ireko ningyo* or *Matryoshka*). Kudô (2000) also noted that such correlations represent a continuum between a group of adverbs and modality forms that denote specific modalities (e. g. necessity, expectation, conjecture, possibility).

In terms of corpus linguistics, these relations are observed as probabilities of co-occurrence and are referred to as '(distant) collocations' (Bekeš 2006, Srdanović et al. 2008a, 2009a). In large-scale empirical studies, Srdanović et al. (2008a,b) explored various suppositional adverbs in numerous Japanese language corpora and subcorpora, revealing, among other findings, the significance of their distribution in detecting different types of corpora.

## 3. Data and methodology

### 3.1. The Japanese language corpora "NINJAL"

The Japanese language corpora used in this study are listed and briefly described in Table 1. They are usually referred to as "NINJAL corpora" relating to the National Institute for Japanese Language and Linguistics (NINJAL), where most of them were developed. The author categorizes these corpora based on their intended use, content, and annotation policies. The abbreviations W and S denote whether the corpus consists of written or spoken language data. The size of the corpora, measured in words, is based on the SUW (short-unit-word) analysis, a method for morphological analyzing lexical units within Japanese language corpora, as outlined in Den et al. (2008).

Two large-scale corpora, BCCWJ and CEJC, prioritize the design of content variability and proportional representation of various linguistic features. They aim to provide a well-balanced representation of contemporary written and spoken Japanese, respectively; therefore, the author classifies them as "balanced corpora".[1]

[1] Balance in terms of corpora is not an absolute but a relative notion. This means that corpus creators strive to create the best representation of overall language patterns, balancing proportions through composition and selection criteria. However, there is no guaranteed method to fully ensure this for the world-wide users of corpora. This aligns with the phenomenon of language change and diversity, which large-scale corpus creators aim to capture and monitor alongside advancements in computational methods.

Table 1: Japanese language corpora used in the analysis (W=written; S=spoken)

| Name<br>(Reference) | W/S | Short description/Keywords | Size |
|---|---|---|---|
| Balanced corpora of contemporary language | | | |
| BCCWJ (Balanced Corpus of Contemporary Written Japanese)<br>(Maekawa et al. 2014) | W | Sampling data from 1976 – 2006; variaty of genres: books, magazines, newspapers, white papers, blogs, etc. | 104,9 million words |
| CEJC (Corpus of Everyday Japanese Conversation) (Koiso et al. 2018) | S | Audio and video data of a variety of natural conversations in daily activities | 2,4 million words – 200h of speech, 577 conversations, 1675 conversants |
| Historical corpora | | | |
| CHJ (Corpus of Historical Japanese)<br>(Kondo et al. 2012) | W | Materials from the following periods: Nara (710–794), Heian (794–1185), Kamakura (1185–1333), Muromachi (1336 to 1573), Edo (1603–1868), Meiji (1868–1912), Taishō (1912-1926), Shōwa (1926-1989), Heisei (1989–2019) | 20,9 million words |
| SSC (Shōwa Speech Corpus) (Maruyama 2020) | S | Materials from 1950 till 1970 in the Shōwa period | 44h of speeches<br>530 000 words |
| SHC (Shōwa -Heisei Corpus of Written Japanese)<br>(Ogiso et al. 2024) | W | Shōwa and Heisei periods, 11 years from 1933 till 2013 | 33,4 million words |
| Specialized language corpora | | | |
| CSJ (Corpus of Spontaneous Japanese)<br>(Maekawa 2003) | S | Mainly monologs of academic presentation speeches; Rich phonetic annotation | 7,5 million words |
| NUCC (Nagoya University Conversation Corpus)<br>(Fujimura et al. 2012) | S | Conversations; Uncontrolled data; mainly informal and standard language but some formal conversations and some dialects included | 100h speeches,<br>1,135 million words |
| COJADS (Corpus of Japanese Dialects)<br>(Kibe et al. 2018) | S | Dialects, all 47 Japanese prefectures, 200 locations | 1,2 million words |

| | | | |
|---|---|---|---|
| CWPC (Gen-Nichi-Ken Corpus of Workplace Conversation) (Kashino et al. 2018) | S | Woman (19) vs. man (21) speeches; various workplaces collected during 1993 and 1999-2000 | 187 000 words |
| Learners' corpora | | | |
| C-JAS (Corpus of Japanese As Second Language) (Sakoda et al. 2014) | S | Longitudinal (3 years), Korean and Chinese native speakers | 46.5h of conversations,<br>570 000 words |
| I-JAS (International Corpus of Japanese as a Second Language) (Sakoda et al. 2016) | W/S | Approx. 1000 learners of 12 different mother tongues; various data covered – oral and written tasks | 3,4 million words |

All of the corpora are incorporated into the skewer-searching system *Kotonoha*. As all of the corpora differ in size, that is, the overall number of lexical units in each corpus, the adjusted frequency PMW (per million words)[2] is used in the analysis. In addition, the author used relative frequency of specific adverbs out of the cumulative number of frequencies of all adverbs within a specific corpus to compare the results and draw conclusions about the data. As the tool provides links to corpus examples, these are occasionally used to check for specifics of usages, such as different prefectures in COJAD, different registers in BCCWJ etc.

### 3.2. Target suppositional adverbs

Twenty-two different suppositional adverbs listed in Table 2 are searched in the corpora and their tendencies and distributions are compared through various categories. Modality types are set based on Kudô (2000), Bekeš (2006) and Srdanović (2008a,b, 2009a,b). The results are summarized using PMW values in

[2] PMW (Per Million Words) calculates the overall corpus size and frequency of specific words, adjusting both values to one million words. This normalization allows corpora of different sizes to be comparable in terms of frequency. The total number of words in each corpus is scaled to one million words, while maintaining the correct ratio between the total number of words and the frequency of the searched word. For example, in a small corpus of 100,000 words, both the corpus size and the frequency of the searched word are multiplied by 10. If a word appears 5 times in the actual corpus, its PMW calculation would show it as 50. Conversely, in a large corpus of 1 billion words, the overall frequency is divided by 1000 to adjust to one million words, and the specific word's frequency is also divided by 1000. Thus, if a word appears 45,000 times in the original corpus, its PMW value would be 45 after adjustment. For further details, refer to the corpus query system "Kotonoha" page.

a table and described in the following section. The most frequent adverb *tabun* and widely used *osoraku* are taken as examples and presented in detail using bar graphs and pie charts provided by the skewer-searching system *Kotonoha*.

The search was set to retrieve specific lemmas with the specified part of speech category "Adverb" (jap. 副詞 *fukushi*), although adjustments were made in some cases due to differences in categorization resulting from the narrow-based approach in Japanese language morphological analysis used for corpus segmentation. The provided meanings are based on the first translation equivalent related to the suppositional sense in the *Goo* dictionary server (or *jisho.org* when not available).

Table 2: Target suppositional adverbs

| | | **Modality** | |
|---|---|---|---|
| **Adverb** | **Transcription** | **type** | **Meaning** |
| 案外・あんがい | angai | POSS | unexpectedly |
| 大方・おおかた | ōkata | EXP | probably |
| 恐らく・おそらく | osoraku | EXP | probably |
| 必ず・かならず | kanarazu | NEC/EXP | certainly |
| 必ずしも・かならずしも | kanarazushimo | NEC | not necessarily |
| ことによると | koto ni yoruto | POSS | perhaps |
| ことによれば | koto ni yoreba | POSS | possibly |
| 嘸・さぞ | sazo | EXP | surely |
| 絶対・ぜったい | zettai | NEC/EXP | absolutely |
| 絶対に・ぜったいに | zettaini | NEC | certainly |
| 大概・たいがい | taigai | NEC/EXP | probably |
| 大抵・たいてい | taitei | EXP | probably |
| 多分・たぶん | tabun | EXP | probably |
| どうやら | dōyara | CON | looks like |
| 急度・きっと | kitto | EXP/NEC | surely |
| どうも | dōmo | CON | somehow |
| ひょっと為たら・ひょっとしたら | hyottoshitara | POSS | possibly |
| ひょっと為ると・ひょっとすると | hyottosuruto | POSS | possibly |
| 若しか為したら・もしかしたら | moshikashitara | POSS | perhaps |
| 若しか為れば・もしかすれば | moshikasureba | POSS | perhaps |
| 若しか為ると・もしかすると | moshikasuruto | POSS | perhaps |
| 余程・よほど・よっぽど | yohodo/yoppodo | CON | quite |

# 4. Distribution differences of suppositional adverbs in various Japanese language corpora

## 4.1. Distribution of 22 suppositional adverbs in 11 corpora

This section presents results of the analysis of the distribution of 22 suppositional adverbs within various Japanese language corpora, extracted using the skewer-searching system *Kotonoha* and then summarized in PMW frequencies for all target adverbs in Table 3. The values are compared and interpreted based on the retrieved data of distribution across all available categories within the system: a) all corpora, b) written and spoken corpora, c) historical periods. The average PMW (Per Million Words) is calculated for each adverb to indicate usages above and below average. The cells with values above average are highlighted in pink. The highest values, indicating the most specific tendencies for each adverb (values up to 300), are highlighted in bold and grey in the table. Values between 100 and 299 are highlighted in grey and italicized.

Table 3: Summary of adverb distribution in the corpora (PWM values)

| Sum of PMW<br>Corpus<br>Adverb | BCCWJ | CEJC | CHJ | C-JAS | COJADS | CSJ | CWPC | I-JAS | NUCC | SHC | SSC | Grand Total | Average PMW |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| tabun | 75 | **1195** | 18 | **1465** | 10 |  | **465** | **2303** | **878** | 49 | 79 | 6759 | 614 |
| zettai | 62 | **386** | 20 | **448** | 50 | 100 | *273* | 71 | **469** | 77 |  | 2073 | 188 |
| dōmo | 53 |  | 89 | 15 | *281* |  | **407** | 51 |  | 74 | **484** | 1778 | 162 |
| kanarazu |  | 74 | *289* | *159* | *133* | *130* | *139* | 48 | 77 |  |  | 1382 | 126 |
| kitto | 73 | *205* | 61 | 22 | 36 | 99 | *107* | 37 | *259* | 51 | *114* | 1063 | 97 |
| osoraku | 75 | 16 | 63 | 0 | 41 | 63 | 59 | 6 | 17 | *113* | *108* | 561 | 51 |
| taitei | 30 | 17 | 70 | 7 | 89 | 17 | 32 | 31 | 18 | 37 | 81 | 431 | 39 |
| zettaini | 52 | 26 | 12 | 26 | 28 | 31 | 21 | 14 | 49 | 64 | 42 | 366 | 33 |
| yohodo/<br>yoppodo | 23 | 27 | 74 | 22 | 49 | 9 | 16 | 1 | 44 | 36 | 55 | 355 | 32 |
| kanarazushimo | 35 | 2 | 80 | 0 | 3 | 37 | 11 | 0 | 8 | 51 | 38 | 266 | 24 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| moshikashitara | 16 | 50 | 1 | 7 | 0 | 32 | 37 | 13 | 41 | 8 | 4 | 209 | 19 |
| taigai | 4 | 3 | 23 | 0 | 57 | 10 | 16 | 0 | 16 | 8 | 68 | 205 | 19 |
| angai | 9 | 14 | 8 | 0 | 9 | 4 | 11 | 1 | 27 | 18 | 47 | 149 | 14 |
| dōyara | 27 | 2 | 14 | 0 | 9 | 9 | 0 | 1 | 13 | 29 | 13 | 118 | 11 |
| ōkata | 2 | 0.8 | 30 | 0 | 9 | 2 | 0 | 0 | 0 | 3 | 2 | 49 | 4 |
| moshikasuruto | 6 | 4 | 0.5 | 0 | 0 | 6 | 16 | 0 | 4 | 5 | 0 | 41 | 4 |
| sazo | 4 | 0 | 24 | 0 | 5 | 0 | 0 | 0 | 0 | 7 | 0 | 40 | 4 |
| hyotoshitara | 5 | 3 | 0.2 | 0 | 2 | 9 | 11 | 0 | 4 | 4 | 0 | 38 | 3 |
| hyottosuruto | 4 | 0.4 | 0.5 | 0 | 3 | 2 | 0 | 0 | 0 | 3 | 0 | 13 | 1.2 |
| kotoniyoruto | 1.3 | 0 | 1.3 | 0 | 0 | 0.8 | 0 | 0 | 0 | 2 | 0 | 5 | 0.5 |
| kotoniyoreba | 0.1 | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| moshikasureba | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| Grand Total | 665 | 2137 | 879 | 2171 | 814 | 889 | 1621 | 2578 | 2031 | 760 | 1358 | 15903 | 1446 |

Based on the retrieved results, the following major findings have been identified:

i) The most widely distributed suppositional adverbs within Japanese language corpora are: *tabun* (6759), *zettai* (2073), *kanarazu* (1382), *kitto* (1063), and *osoraku* (561). *Dōmo*, although highly frequent, is not highlighted here as it appears in various phrases without functioning as a suppositional adverb. On the other hand, the least represented include: *moshikasureba* (0), *kotoniyoreba* (0), *kotoniyoruto* (0.5).

ii) *Tabun* (6759) is predominantly used in conversational settings, particularly informal ones (1195 in CEJC, 878 in NUCC, 465 in CWPS), and it is notably overused by learners of Japanese as a foreign language (I-JAS, C-JAS). Refer to section 4.2 for more details..

iii) *Osoraku* (561) is predominantly used in written Japanese, in formal settings, with significant usage in historical data (113 in SHC, 108 in SSC), especially in the materials from the Shōwa period. It ranks as the sixth most frequent suppositional adverb overall but registers zero frequency in the learners' corpus C-JAS and remarcably low occurrence in I-JAS, indicating that it is underused by Japanese language learners. Refer to section 4.2 for more details..

iv) Suppositional adverbs are predominantly represented in spoken corpora (e.g., *tabun*, *zettai*, *dōmo*, *kitto* in CEJC, CWPS, NUCC, SSC), reflecting their significant roles in spoken communication. However, some of these adverbs appear more frequently in written data (*osoraku*, *kanarazushimo*, *dōyara*, *moshikasuruto* in SHC, CHJ, BCCWJ) or formal spoken contexts (*osoraku* in CSJ and SSC). The total PMF frequency of these adverbs in each spoken corpus, including learners' data, exceeds the average total PMW frequency of these adverbs across all corpora (e.g., grand total 2137 in CEJC vs. average 1446), further confirming their importance in spoken interaction..

v) Some of the targeted adverbs are specifically noted for historical usage, predominantly in the Shōwa period but also in earlier periods (*yoppodo*/*yohodo*, *taigai*, *ōkata*, *kanarazushimo*, *sazo*, along with *kanarazu* which is despersed across various periods including contemporary texts covered in BCCWJ).

vi) The I-JAS learners' corpus appears to be highly specialized in terms of the types of data collected, involving specific tasks, while C-JAS is smaller and consists of conversational data from six learners. I-JAS exhibits notably high usage of *tabun*, with all other adverbs being comparatively underrepresented compared to native speakers. In contrast, C-JAS shows frequent usage of *zettai* and *kanarazu* among learners. More than half of the adverbs are rarely or never used by learners (11 adverbs in C-JAS and 6 in I-JAS have no occurrences), despite learners' overall usage of a limited number of adverbs being above the average usage found across all corpora.

vii) The least frequent adverbs (*sazo, hyottoshitara, hyottosuruto, kotoniyoruto, kotoniyoreba, moshikasureba*) are predominantly used in written corpora rather than in spoken conversations. Specifically, they appear more frequently in corpora like BCCWJ and SHC, which are noted for their balanced coverage of various adverbs, including the rare ones.

viii) Regarding modality types, expectation and necessity are the most widely represented, whereas conjecture and possibility are less frequently used. This observation aligns with findings from Srdanović (2009a,b), which analyzed a different set of Japanese language data.

### 4.2. The adverbs *tabun* and *osoraku* ‘probably’ in the corpora

Table 4 displays distribution of the suppositional adverbs *tabun* and *osoraku*, both of which convey the meaning ‘probably’ and fall under the modality type of expectation. It provides the actual frequencies of these adverbs within specific corpora, along with the overall corpus data.

Table 4: The adverbs *tabun* and *osoraku* in Japanese language corpora (actual frequency)

| Corpus | Tabun freq. | Freq. of suw (overall) | Osoraku freq. | Freq. of suw (overall) |
|---|---|---|---|---|
| BCCWJ | 7,899 | 104,911,460 | 7,846 | 104,911,460 |
| CSJ | 1,681 | 7,576,046 | 481 | 7,576,046 |
| CEJC | 2,890 | 2,419,171 | 38 | 2,419,171 |
| SSC | 42 | 528,589 | 57 | 528,589 |
| NUCC | 997 | 1,135,329 | 19 | 1,135,329 |
| CWPC | 87 | 186,906 | 11 | 186,906 |
| CHJ | 332 | 18,650,926 | 1,179 | 18,650,926 |
| SHC | 1,622 | 33,404,844 | 3,791 | 33,404,844 |
| COJADS | 12 | 1,221,624 | 50 | 1,221,624 |
| C-JAS | 396 | 270,333 | 0 | 270,333 |
| I-JAS | 7,836 | 3,401,933 | 21 | 3,401,933 |

Normalized and compared frequencies, considering the overall size of each corpus, are displayed in Figure 1 and Figure 2. The differences in distribution and usage of these two adverbs with similar meanings are apparent from the results presented in the table and figures. *Tabun* is predominantly used in spoken corpora (CEJC, NUCC, CWPC) rather than in written form, although its actual frequency in BCCWJ is also considerable. The adverb *tabun* appears more frequently in informal everyday conversations than in formal (academic presentations) settings (CEJC vs. CSJ). In contrast, *osoraku* is, unlike *tabun,* used more in written or formal spoken contexts (CHJ, SHC, BCCWJ; SSC, CEJC).

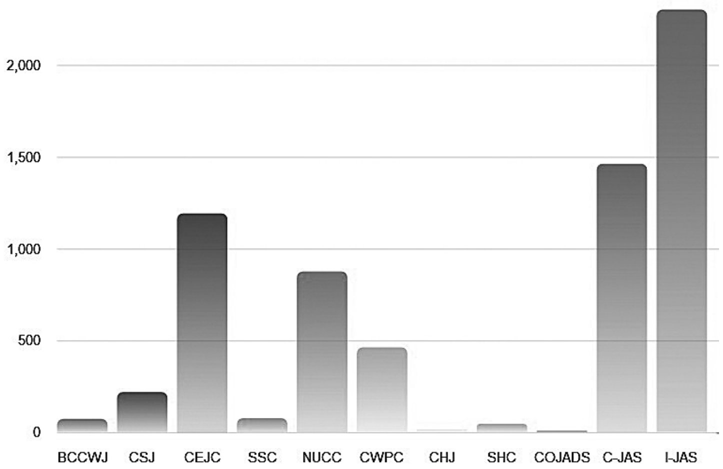

Figure 1: The adverb *tabun* in various Japanese language corpora (PMW normalized frequency: 2000 scale)

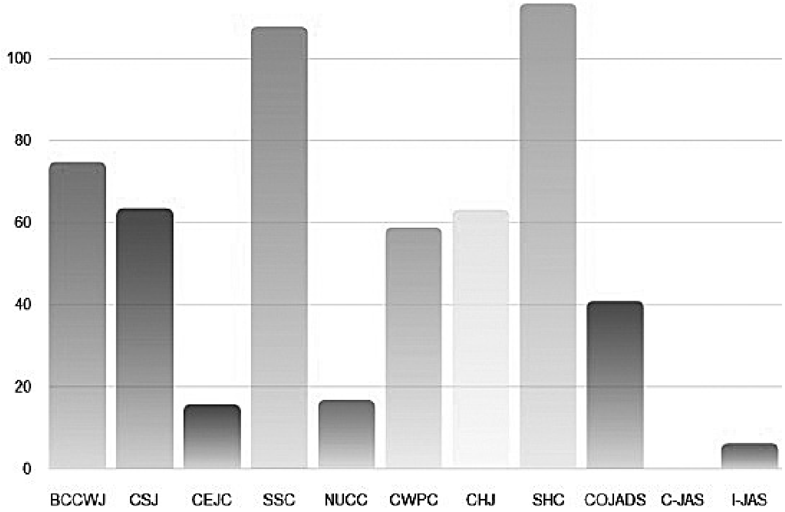

Figure 2: The adverb *osoraku* in various Japanese language corpora (PMW normalized frequency: 100 scale)

Obviously, *osoraku* is predominantly used in formal settings, most notably in CSJ, which consists mainly of academic speeches, and CWPC, which generally features more formal contexts compared to other spoken corpora. Additionally, the extensive usage of *osoraku* in historical corpora (CHJ, SHC, SSC) suggests the need for exploring its usage from a diachronic perspective, which will be elaborated below.

Figure 3 summarizes the usage of these two adverbs in pie charts depicting their distribution across written and spoken corpora. The usage of *tabun* is illustrated on the left side, while the usage of *osoraku* is shown on the right. Although this chart does not differentiate between formal and informal settings, which would also provide valuable insights into the differences in their usage, it clearly highlights their tendencies in distribution between written and spoken contexts.

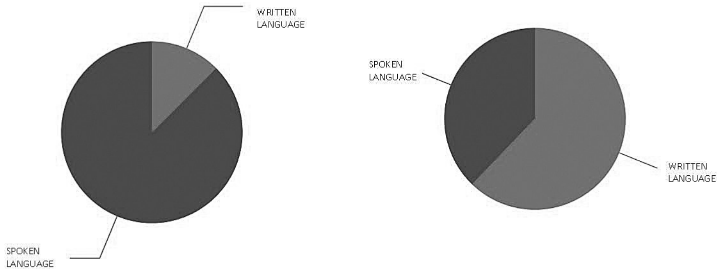

Figure 3 Adverbs *tabun* (left-side) and *osoraku* (right-side) in written and spoken corpora

Interestingly, the learner's corpora (especially I-JAS, but also C-JAS) show a significantly higher occurrence of the adverb *tabun* compared to native speakers' data (Figure 1). This suggests that learners of Japanese as a foreign language tend to overuse the adverb *tabun*. In contrast to *tabun*, *osoraku* is underrepresented in learners' data, with zero occurrences in C-JAS and very low frequencies in I-JAS (Figure 2). Possible interpretations include: i) students may overuse the adverb *tabun* because they learn it before the other adverbs with similar meaning; ii) learners' corpora may contain more informal data due to the task settings or language usage styles of learners; iii) learners may communicate with greater uncertainty in their language skills, relying more on suppositional adverbs like *tabun* as a communication strategy while preparing to produce an utterance or written content; iv) *osoraku* is not sufficiently represented in language textbooks.

It should be noted that normalized frequencies presented in Figure 1 and 2 require careful interpration due to differences in scale (2000 vs. 100). This is particularly significant for the adverbs *tabun* and *osoraku*, as they have similar frequencies in BCCWJ, but their values are presented in different scales due to their distinct usage patterns in specific corpora. In Figure 1, the bar for BCCWJ appears relatively low on the scale due to exceptionally high values for *tabun* in I-JAS, whereas in Figure 2, the scale is lower and the bar for BCCWJ is comparatively higher for *osoraku*. Therefore, when interpreting the data, attention to the scale values is crucial. As *tabun* and *osoraku* have similar frequencies in BCCWJ, examining subcorpora revealed that *tabun* is predominantly used in informal types of texts such as blogs and Yahoo! Chiebukuro forums, whereas *osoraku* appears more frequently in formalized texts such as National Diet meeting minutes, books, and textbooks. This distinction is also noted by Maebo (2012).

From a diachronic perspective, it can be observed that the adverb *tabun* did not appear in the available materials from the Nara, Heian, and Kamakura periods. It first emerged in materials from the Muromachi period, with fewer occurrences in data from the Edo period. Its usage then gradually increased until the Heisei period, during which it became the most widely used, based on available corpus data.

As mentioned earlier, the adverb *osoraku* shows significant occurrences in SSC and SHC, indicating a slight decrease in contemporary usage compared to the Shōwa period. This trend was confirmed by analyzing corpora using diachronic criteria, as illustrated in Figure 5. The results demonstrate that *osoraku* is most prevalent in materials from the Shōwa period, with above-average occurrences also noted in the Meiji and Heisei periods. Its earliest usages are recorded from the Kamakura period.

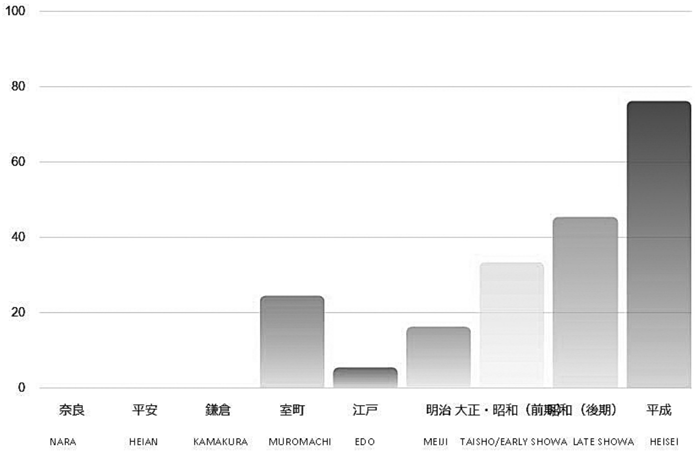

Figure 4 Usage of adverb *tabun* in data from specific historical periods

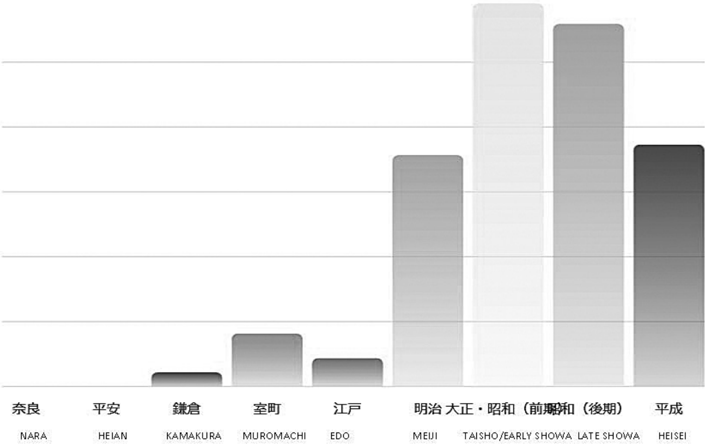

Figure 5 Usage of adverb *osoraku* in data from specific historical periods

From a historical standpoint, although important trends in the usage changes of lexical units are revealed, caution must be exercised in drawing conclusions. These changes may still primarily reflect the characteristics of the corpus and its register rather than the broader language usage within the specific historical period.

## 5. Remarks for further development

The analysis revealed opportunities for enhancement of the resources used, particularly in terms of morphological analysis of corpora and the further development of the concordance tool.

Some adverbs are not recognized as single units but rather morphologically analyzed into smaller components, which can complicate searches but remain feasible with specific search algorithms. It is desirable that these lexical units are morphologically analyzed and searchable as single-word units as well. In cases where such elements are calculated together with another word (e.g., *kanarazu* as part of *kanarazu+shimo*), manual checking and removal are necessary. Regarding results for *dōmo*, it should be noted that it frequently appears within various phrases.

There are still cases when it is difficult to differentiate between part-of-speech categories because units with different function and categories are morphologically analyzed in the same way. For example, *zettai* and *taigai* are categorized as *noun-general-adverbial* (名詞-普通名詞-副詞可能, *meishi-futsū_meishi-fukushi_kanō*), which includes nouns such as 絶対数 *zettai-sū* 'apsolute value' or 絶対性 *zettai-sei* 'absoluteness', along with adverb. Automatic differentiation between such items would be desirable. Because of automatic morphological analysis and its narrow approach, methodologically, it is always advisable to conduct random sample searches to qualitatively analyze a specific subset of data.

The skewer-searching system allows for searching examples of use within any of the corpora. Enhancements in further retrieving examples based on specified categories per subcorpus and using PMW would be beneficial, along with differentiating between formal and informal data.

## 6. Conclusion and further work

This research provided the distribution of the targeted suppositional adverbs across eleven different Japanese language corpora, using the novel state-of-the-art concordance tool for skewer-searching language data, *matomete kensaku “KOTONOHA”*. Twenty-two adverbs expressing various modalities—expectation, necessity, possibility, and conjecture—are explored in terms of their empirical usage in: a) written and spoken corpora, b) genres and registers, c) native speakers’ and learners’ corpora, and d) historical periods in diachronic corpora. Two of these adverbs, *tabun* and *osoraku*, are presented in detail with accompanying charts, graphs, and tables.

The search tool, corpora, and applied methodology reveal usage tendencies of the examined items across specified categories, with major findings summarized in sections 4.1 and 4.2. The most frequent suppositional adverbs appear in spoken corpora, highlighting the importance of these adverbs in human interaction. Written corpora also exhibit typical suppositional adverb usage. Adverb usage tends to vary based on genre and register, as exemplified by the differences in uses of *tabun* and *osoraku*. Expectation type of suppositional adverb modality, followed by necessity, is most prominent according to the empirical analysis. Significant differences in adverb usage between native speakers and learners have been discovered and need further detailed examination. Additionally, the diachronic usage of adverbs sheds light on language changes, revealing both archaic usages and a dynamic picture of adverb usage over time. Beyond the detailed examination of adverb usage and distribution, the analysis also portrays and elucidates corpus types, their typical genres, and characteristics.

The remarkable advancements in developing corpora and other resources for the Japanese language provide great opportunities for linguistic studies across diverse aspects, including language education. However, interpreting the data entails potential risks, requiring a thorough examination of results from multiple perspectives, encompassing various categories and values, as well as exact and normalized frequencies.

This analysis focuses on the suppositional adverbs specified in Kudô’s (2000), Bekeš’s (2006), and Srdanović’s (2008a,b, 2009a,b) work. Future research will explore additional adverbs and examine their tendencies in expressing suppositional modality. A detailed analysis of adverb usage in specific contexts and settings, with provided examples, is planned to uncover more about their multifunctional roles in both spoken and written language, facilitated by compiled corpora and advanced analytical tools.

The research findings contribute to a better understanding of linguistic features and the usage of adverbs, their dependence on context, language users, and language evolution. Furthermore, they have implications for Japanese language education and offer suggestions for future research on this topic. The paper also presents the methodology and framework for future empirical studies on Japanese language data using multiple corpora and criteria. It provides recommendations for further advancement of resources and introduces the latest developments in Japanese language corpora and tools, which serve as a potential stimulus for future research on other languages and datasets.

## Acknowledgments

The work is supported by the Japan Foundation under the Japanese Fellowship Grant 10148587 and project title “Distant collocations in spoken and written Japanese corpora: the case of adverbs and modality forms (日本語書き言葉・話し言葉コーパスにおける遠隔共起 ― 副詞とモダリティ形式を例として ― )”. This research is related to the institutional project “Japanese language and society: basic concepts and empirical analysis (Japanski jezik i društvo: temeljni pojmovi i empirijska analiza)” conducted at the Juraj Dobrila University of Pula Faculty of Humanities Department of Asian Studies (leader: Prof. Irena Srdanović, period: 2023-2028). The author of the paper collaborates on related projects at the National Institute of Japanese Language and Linguistics.

# References

Bekeš, Andrej. 2006. Japanese suppositional adverbs in speaker-hearer interaction. *The third conference on Japanese language and Japanese language teaching: Proceedings of the conference*. Ed. Tollini, A. Libreria editrice cafoscarina. 34–48.

Den, Yasuharu et al. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* 26(1-2). 129-148.

Fujimura, Itsuko; Chiba, Shoju; Ohso, Mieko. 2012. Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*, 393–398.

Ishiguro, Kei. 2023. *Komyuryoku wa "fukushi" de kimaru*. Kabushiki kaisha kōbunsha. Tokyo.

Joyce, Terry; Hodošček, Bor; Nishina, Kikuko. 2012. Orthographic representation and variation within the Japanese writing system: Some corpus-based observations. *Written Language and Literacy* 15. 254–278. doi.org/10.1075/wll.15.2.07joy.

Kashino, Wakiko et al. 2018. Supplemental arrangement for public data available in the Chunagon versions of 'Gen-Nichi-Ken corpus of workplace conversation'. *Proceedings of Language Resources Workshop 2018*. 494–509.

Kibe, Nobuko; Otsuki, Tomoyo; Sato, Kumiko. 2018. Intonational variations at the end of interrogative sentences in Japanese dialects: From the "Corpus of Japanese Dialects". *Proceedings of the LREC 2018 Special Speech Sessions*. 21–28.

Kudô, Hiroshi. 2000. Fukushi to bun no chinjutsu no taipu (Adverbs and the type of clause-final modality). *Nihongo no bunpō 3 – Modariti (Japanese grammar 3: Modality)* Eds. Nitta, Y.: Masuoka, T. Iwanami shoten. Tokyo. 161–234.

Koiso, Hanae et al. 2018. Construction of the corpus of everyday Japanese conversation: An interim report. *Proceedings of the 11th edition of the Langauge Resources and Evaluation Conference*. 4259–4264.

Kondo, Yasuhiro. 2012. The NINJAL diachronic corpus project - Oxford VSARPJ project joint symposium corpus based studies of Japanese language history. *NINJAL Project Review* 3. 84–92.

Ogiso, Toshinobu et al. 2024. Design, Construction and Publication of the Shōwa - Heisei Corpus of Written Japanese. *Jōhō shori gakkai ronbunshū* 65/2. 278-291. doi.org/10.20729/00232292.

Oka, Teruaki et al. 2020. KOTONOHA: A Corpus Concordance System for Skewer-Searching NINJAL Corpora. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 7077–7083. https://aclanthology.org/2020.lrec-1.875 (accessed 5 July 2024).

Maebo, Kanako. 2012. Kōpasu ni okeru *tabun osoraku* no shiyō keikō no bunseki. *The Hitotsubashi Journal for Japanese language* 1. 49-60. https://ndlsearch.ndl.go.jp/books/R000000004-I031679196 (accessed 5 July 2024).

Maekawa, Kikuo. 2003. Corpus of spontaneous Japanese: its design and evaluation. *Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition,* paper MMO2. https://www.researchgate.net/publication/228584595_Corpus_of_Spontaneous_Japanese_Its_design_and_evaluation (accessed 5 July 2024).

Maekawa, Kikuo et al. 2014. Balanced corpus of contemporary written Japanese. *Language Resourses and Evaluation* 48/2. 345–371. doi.org/10.1007/s10579-013-9261-0.

Maruyama, Takehiko. 2020. On the Possibility of a Diachronic Speech Corpus of Japanese. *Japanese Language from Empirical Perspective: Corpus-based studies and stud-*

*ies on discourse*. 219-234. Eds. Srdanović, Irena; Bekeš, Andrej. Znanstvena založba FF. Ljubljana. https://issuu.com/znanstvenazalozbaff/docs/the_japanese (accessed 5 July 2024).

Masuoka, Takashi; Takubo, Yukinori. 1989. *Kiso nihongo bunpō* (Basic Japanese Grammar). Kuroshio. Tokyo.

Minami, Fujio. 1974. *Gendai nihongo no kōzō (Structure of the Modern Japanese language).*Taishūkan shoten. Tokyo.

Sakoda, Kumiko et al. 2014. C-JAS (Corpus of Japanese as a second language) kōchiku ni kansuru hōkoku-sho. *Daigaku kyōdō riyō kikan hōjin ningen bunka kenkyū kikō.*

Sakoda, Kumiko et al. 2016. International corpus of Japanese as a second language. *NINJAL Project Review* 6. 93–110.

Srdanović, Irena; Bekeš, Andrej; Nishina, Kikuko. 2008a. Distant Collocations between Suppositional Adverbs and Clause-Final Modality Forms in Japanese Language Corpora. *Large-Scale Knowledge Resources. Construction and Application. LKR 2008*. Eds. Tokunaga, T.; Ortega, A. *Lecture Notes in Computer Science* 4938. Springer. Berlin – Heidelberg.

Srdanović, Irena; Bekeš, Andrej; Nishina, Kikuko. 2008b. Fukusū no kōpasu ni mirareru fukushi to bunmatsu modariti no enkaku kyōki kankei. *Proceedings of Nihongo kōpasu Heisei 19 nen-do kōkai waakushoppu (yokōshū)*. 223-230.

Srdanović, Irena; Bekeš, Andrej; Nishina, Kikuko. 2009a. Kōpasu ni motozuita goi shirabasu sakusei ni mukete (Towards corpus-based creation of lexical syllabus). *Nihongo kyōiku (Journal of Japanese Language Teaching)* 142. 69–79.

Srdanović, Irena et al. 2009b. Extraction of Suppositional Adverb and Clause-Final Modality Form Distant Collocations Using a Web Corpus and Corpus Query System and its Application to Japanese Language Learning. *Journal of Natural Language Processing* 16. 29–46. doi.org/10.5715/jnlp.16.4_29.

Yamada, Yoshio. 1936. *Nihon bunpō gaku gairon (Survey of Japanese Grammar).* Hōbun kan. Tokyo.

## Sources from the Internet

*Corpus catalog* (NINJAL). https://clrd.ninjal.ac.jp/en/corpus-list.html

*Matomete kensaku* “KOTONOHA” まとめて検索『KOTONOHA』. https://chunagon.ninjal.ac.jp/integrated/

*Goo dictionary server.* https://dictionary.goo.ne.jp/

*Jisho dictionary*. https://jisho.org/

# Distribucija priloga pretpostavke u korpusima japanskog jezika: korištenje sustava za pretraživanje KOTONOHA

## *Sažetak*

Ovaj rad ima za cilj analizirati i opisati distribuciju modalnih priloga u korpusima japanskog jezika koristeći najnovije resurse za empirijsku jezičnu analizu. Specifični prilozi s modalnom funkcijom kao što su *tabun*, *osoraku*, *kanarazu*, *kanarazushimo*, *kitto* i drugi, uspoređuju se u njihovoj upotrebi unutar višestrukih korpusa uz korištenje sustava pretraživanja „KOTONOHA", razvijenog na Nacionalnom institutu za japanski jezik i lingvistiku. Analiza otkriva koji su prilozi češće korišteni u različitim kontekstima, a koji su specijaliziraniji. Uspoređuje se upotreba priloga među dvanaest različitih korpusa japanskog jezika, uključujući pisane i govorne korpuse, korpuse učenika, dijakronijske korpuse, što omogućuje uvid u razlike u žanrovima, registrima, te u upotrebi od strane izvornih govornika ili učenika, kao i promjene kroz povijesne periode. Dva često korištena priloga, *tabun* i *osoraku* 'vjerojatno, možda', koji pripadaju tipu modalnosti očekivanja, koriste se kao model za prikaz prikupljenih podataka i grafova s detaljnom interpretacijom. Istraživanje doprinosi boljem razumijevanju jezičnih značajka priloga, njihove kontekstualne uporabe i razvoja s implikacijama za poučavanje japanskog jezika. Također predlaže metodologiju za buduća istraživanja korištenjem više korpusa i sugerira napredak u jezičnim resursima kako bi se potaknula daljnja istraživanja.

***Keywords:*** modal adverbs, Japanese language, distribution, language corpora, suppositional adverbs

***Ključne riječi:*** modalni prilozi, japanski jezik, distribucija, jezični korpusi, prilozi pretpostavke