Weaving a Dictionary with AI

サンプルID	開始位置	連番	前文脈	キー	後文脈	語彙素読み	語彙素	語彙素細分類	品詞	活用型	活用形	レジスター	執筆者	書名/出典	編著者等	出版者	出版年	話者名
PB29_00256	44140	28520	を|言おうとした|。# だ|が|彼女は|私|を抱擁した|。# 彼女|の|体|に|	手	|をまわすと|、|肩押|骨|に|手|が|触れた|。# 彼女は|さらに|強く|私|を抱きしめた	テ	手		名詞-普通名詞-助数詞可能			出版・書籍	レナード・チャン(著)/三川 基好(訳)	夜明けの挽歌	レナード・チャン|著;三川基好|訳	アーティストハウス;角川書店(発売)	2002	
LBo0_00018	58580	35580	まま|稀少性|という|こと|に|は|ならない|。# 「ボク|が|ようやく|『|オカルト|物語|』|を|	手	|に|入れた|の|は|それから|約|一|年|後|のこと|。# 見つけた|場所|は、|祐天寺	テ	手		名詞-普通名詞-助数詞可能			図書館・書籍	牧 眞司(著)	ブックハンターの冒険	牧眞司|著	学陽書房	2000	
PB59_00165	9180	5560	て|も、|統は|持た|なかっ|た|。# 自衛|の|ため|に|東洋|の|武術|を習った|。#	手	|や|足|を|使い、|狙い|を|定めて|人|を殺すこと|も|できる|格闘技|だ|。#	テ	手		名詞-普通名詞-助数詞可能			出版・書籍	ヘレン・ビアンチン(著)/鈴木 けい(訳)	甘い屈辱	ヘレン・ビアンチン|作;鈴木けい|訳	ハーレクイン	2005	
PB29_00679	83790	55250	、|今まで|あの|子は|海賊|の|仲|に|なって|た|わけ|です|から|ね|。# 短剣|が|	手	|に|戻った|今|ぢゃ|あ、|あの|子は|海賊|を|殺す|こと|に|なる|ん|です|。# 重政	テ	手		名詞-普通名詞-助数詞可能			出版・書籍	三島 由紀夫(著)	三島由紀夫全集	三島由紀夫|著	新潮社	2002	

[What are the Japanese corpora?]
The Japanese corpora are publicly accessible linguistic databases that have collected large amounts of both written and spoken Japanese. The National Institute for Japanese Language and Linguistics (NINJAL) has taken the lead in developing the databases, including the Balanced Corpus of Contemporary Written Japanese (BCCWJ), the Historical Corpus of Japanese (CHJ), the Corpus of Japanese Dialects and the Corpus of Spoken Japanese. BCCWJ contains 104.3 million words drawn from a wide range of sources: books, magazines, newspapers, white papers, blogs, online forums, textbooks, legal texts and more. These are publicly accessible online. Besides the paid offline version, there are two free applications: *Shonagon* (no registration required) and *Chunagon* (registration required). *Shonagon* allows users to search BCCWJ, while *Chunagon* provides access to all corpora. The image above shows a sample search result for the word '*te* (hand)' in BCCWJ using Chunagon.

Professor Emeritus Kondo Yasuhiro (Aoyama Gakuin University) talks about AI technology applied to the revision of the Japanese language dictionary, interviewed by Matsui Mio

Japan's largest and only large-scale dictionary of the Japanese language, the Second Edition of Nihon Kokugo Daijiten (Shogakukan Unabridged Dictionary of the Japanese Language; please refer to box), is undergoing its first major revision in 30 years. Shogakukan, the publisher of the dictionary, has announced that the Third Edition is scheduled to be published in 2032. They have considered the adoption of new digital technologies such as the Japanese corpora (please refer to photo caption) and AI for its editorial work. We spoke with Professor Emeritus Kondo Yasuhiro, an editorial board member and a leading corpus linguist, about how *Nihon Kokugo Daijiten* will shape the future of the dictionary and the Japanese language—and what role AI may play.

[*Nihon Kokugo Daijiten,* Second Edition]
13 volumes + a supplementary volume
Kitahara Yasuo, Kubota Jun, Taniwaki Masachika, Tokugawa Munemasa, Hayashi Oki, Maeda Tomiyoshi, Matsui Shigekazu, Watanabe Minoru
Shogukukan, published from 2000 to 2002
The Second Edition is Japan's largest and only large Japanese dictionary, containing 500,000 entries. It draws upon approximately 30,000 literary and historical sources—from *Kojiki (Records of Ancient Matters)* and *Nihon Shoki* to modern literature in the Showa era—collecting one million usage examples. More than 3,000 experts from a wide range of fields, including Japanese linguistics, classical literature, social sciences, and natural sciences, contributed to its completion. The dictionary precisely captures the changing meanings of words over time, offering a rich reflection of both the Japanese language and Japanese culture as a whole. 60 years since the First Edition was published, the revision of the Third Edition has begun, scheduled for release in 2032 to mark the 110th anniversary of Shogakukan. The announced editorial board members for the new edition include linguists Kinsui Satoshi, Kondo Yasuhiro, Tanaka Makiro, Hidaka Mizuho, Maeda Naoko and Yamamoto Shingo.

[*Nihon Kokugo Daijiten*, Second Edition, a headword, *Sora* (sky)]
Starting with the first appearance of the word in *Kojiki*, the dictionary shows the examples of its use in *Manyoshu*, *Tsurezuregusa (Essays in Idleness)* and more sources from various periods, allowing users to follow the shift in the usage of the word across different eras. It offers a clear and thorough view of the word's historical transition.

The corpus revolutionizes the future of the Japanese language

Sixty years after the First Edition was published in 1972, the editing work of the Third Edition of *Nihon Kokugo Daijiten* (hereinafter called "Nikkoku") has begun, set to be released in 2032. The Second Edition, which is currently on sale, is Japan's largest and only large-scale Japanese language dictionary, with a total of 13 volumes and 500,000 words. The Third Edition is aimed at significant improvement of its contents and the editorial process: collecting older and more up-to-date usage examples, adding 30,000 to 50,000 new entries, and data sharing on the cloud. In addition, it is planned for sequential release on the internet platform JapanKnowledge.

In anticipation of the further evolution and penetration of IT in eight years' time, attempts are being made to introduce AI and other digital technologies to both the editing work and the use of the dictionary. Corpus linguistics is one of the cutting-edge research fields related to AI that is expected to play a central role in the revision of *Nikkoku*. To put it simply, a corpus is a database of words, a large collection of written and spoken texts that can be searched on a computer. Corpus linguistics is the study of building and analyzing corpora. Kondo Yasuhiro (Professor Emeritus, Aoyama Gakuin), an editorial board member of the Third Edition of *Nikkoku*, has been involved in this research since the 1970s, when he entered the Faculty of Letters at the University of Tokyo, and is a prominent academic of corpus linguistics today.

Corpora of the Japanese language are accessible to the public via the internet. The National Institute for Japanese Language and Linguistics (NINJAL) has been taking the lead in constructing the "Balanced Corpus of Contemporary Japanese Written Language (BCCWJ)" and the "Corpus of

Historical Japanese (CHJ)," which are available free of charge on its website (please refer to photo caption). Kondo was the first team leader in the construction of CHJ. This corpus contains an enormous number of texts from the Nara period (710–790) to the Meiji (1868–1912) and Taisho (1912–25) periods, including *Manyoshu*[1], *Genji Monogatari (Tale of Genji)*, *Makura no Soshi* (*The Pillow Book*), *Kokin Waka Shu* (collection of ancient and modern poetry, usually abbreviated to *Kokinshu*)[2], and modern newspapers, magazines and novels. Academics and experts in the history of the Japanese language have utilized CHJ as a fundamental resource for their study.

The corpus reveals true meanings of words

Now, the question is how corpus linguistics can be applied to revise *Nikkoku*. Kondo says, "I am proposing that the whole of the Second Edition of *Nikkoku*, which I was also involved in revising, should be incorporated into a searchable corpus. This would definitely be very beneficial for the revision." Unlike printed dictionaries, which generally restrict searches to headwords, a corpus enables effortless access to the full body of text. "This would help us verify whether each definition properly corresponds to its headword. There are many different headwords that refer to the same concept. Yet, some of their definitions contradict each other. We would be able to identify and correct such inconsistencies by using the corpus. Another point to make sure is that the definitions are not circular. The definitions of Japanese words tend to be circular by nature—like describing '*ishi* (stone)' as 'smaller than a rock,' and '*iwa* (rock)' as 'a large piece of stone.' Still, when the circularity is too extreme, we need to avoid it."

What further role does Kondo envision for corpus linguistics? "Most importantly, I hope that it will contribute to improving, even just a little, the classification of word definitions." Kondo has considered the classification to be a challenging task ever since the revision of the Second Edition conducted 30 years ago. "The word definition has 'branches,' in other words, subdivisions. Looking up '*te* (hand)' in a dictionary, the first branch is 'a part of the body,' and one of the other branches is 'Sono *te* ga attaka (I never thought of that,' with '*te*' here meaning 'method' or 'means'). The thing is we are not quite sure if this classification is right. '*Te*' is fairly easy to classify into branches, but not all words are like that. Using the corpus would likely lead to more precise and reliable classifications."

One of the great benefits of the corpus is that it display the context before and after a word. A search for '*te* (hand)' (using *Chunagon*) in the aforementioned BCCWJ yields 49,567 hits. These include 'shorishi owatte, *te* wo arau (I finished that and washed my hands)' (Tendo Arata, *Eien no Ko* [Eternal Child]) and 'Sekkaku *te* ni shokumo ari (since you have a skill, ...)' (Yahoo! Chiebukuro [wisdom bags]). Kondo explains, "Extracting this many samples manually would be impossible. But with the corpus, we can accumulate and examine vast amounts of linguistic data. It also shows co-occurring words (words that appear together frequently, especially in a specific order). The corpus reveals to us various kinds of information about a word that aren't immediately obvious when we just look at the word. Through this method, we would surely be able to gain deeper insights into the word's authentic meaning." Definitions

[1] Published at the end of the Nara period (710–794), this is the oldest collection of *waka* poetry in Japan, containing some 4,500 poems in 20 volumes by a wide range of poets, from emperors to farmers.

[2] The Imperial Anthology of Waka Poetry, completed in 905, consists of 20 volumes and contains about 1,100 poems. The collection of imperial anthologies of *waka* poetry from *Kokin Waka Shu* to the eighth, *Shin Kokin Waka Shu* (New Collection of Ancient and Modern Poetry), is specifically referred to as *Hachidai Shu* (Eight Great Anthologies of Waka Poetry).

and classifications of words have largely relied on the knowledge and intuition of specialists so far, but the corpus allows for thorough analysis of them with quantitative backing.

AI analysis uncovers the essence of words that is beyond human perception

Kondo's main research is currently focused on more advanced AI analysis of classical Japanese, and he is hoping that it will contribute to improving word definitions in the Third Edition of *Nikkoku*. "AI analysis" sounds like complete gibberish to someone like us without a strong background in science, so we ask Kondo to put it in easy-to-understand language. "AI converts words into numbers. By comparing these numerical values, AI can tell us whether words are similar or not. A good example is an online shopping site. If you want to buy something *mofumofu* (fluffy or woolly) and you enter '*mofumofu*' in the search space, a cashmere sweater might come up. Although the word '*mofumofu*' is not in the sweater, the AI-based search in the site determines that the two are similar. This is because the numerical values that 'mofumofu' and 'cashmere sweater' have are close to each other." This means that AI has enabled a corpus-based semantic analysis. "*Chunagon* cannot be used for this purpose, as it is just a search application. It cannot answer if '*te* (hand)' and '*ashi* (foot)' are similar. But AI can do so by comparing the numerical values of these two words. It can be said that AI has raised corpus linguistics to a higher stage."

What is remarkable about representing words as numbers is that it can detect nuances and meanings imperceptible to humans, offering a way to get closer to the essence of words. One example of this is Kondo's ongoing research into the differentiation in language use based on gender in ancient times. "I aim to find whether there were any words or expressions that were preferentially used by men or women in the Heian period. Modern Japanese is full of them, like '*yone*?' by women and '*daro*?' by men (both mean 'you know'). These kinds of gendered words frequently appear in novels. On the other hand, we don't really know if people in the Heian period had those specific words. We don't share the ancient people's sense of language or their mentality, or what we call 'introspection.' Therefore, even if female or male words appear in classical literature, we cannot recognize them as such. However, if we process a corpus of classical languages using AI, we may be able to find those words." This means that the introspection of the ancients can be acquired via AI analysis. "If we achieve this, we can add an explanation, 'this was a word exclusively used by females/males,' to the definition in the Third Edition of *Nikkoku*."

A new picture of the ancient world may emerge through AI analysis

Kondo says that trends in language can also be extracted through AI analysis. "I am currently researching anthologies of *waka* (classical Japanese poetry) such as *Kokin Waka Shu* and *Ogura Hyakunin Isshu* (the later name for *Hyakunin Isshu,* One Hundred Poems by One Hundred Poets). AI converts each *waka* poem into numbers, and based on that data I analyze themes that those two anthologies focus on." *Kokin Waka Shu* was published in the early Heian period. It was compiled in 905 by Ki no Tsurayuki (872–945) and other respected noble poets at the Emperor Daigo's order. "Almost all *waka* anthologies mainly deal with scenery and emotions. *Kokin Waka Shu* particularly

focuses on 'birds' and 'flowers.' Poets enjoy hearing birds singing and admire the beauty of flowers. As this aligns with findings from previous research, you can see that the computer's calculations are correct." At the same time, it can be said that the AI analysis demonstrates the validity of earlier studies.

"Then, how about *Ogura Hyakunin Isshu*?," says Kondo. This best-known *waka* anthology is thought to have been completed around the early 13th century in the Kamakura period (1185–1333). Fujiwara no Teika (1162–1241) is commonly credited with compiling this, but some scholars have raised alternative theories. "*Kokin Waka Shu* and the *Hyakunin Isshu* look similar but, in fact, have very different characteristics. With a gap of more than 300 years between the publication of the two, it is natural that the themes in the poems have changed. In *Hyakunin Isshu*, words related to 'water' and 'night,' such as 'seaside' and 'riverside,' and 'evening' and 'dusk,' appear very frequently.' These two themes are difficult to find, even for experts. There are only a hundred poems in *Hyakunin Isshu*, so noticing the trends and patterns is rather hard. But a computer can do this through calculations." Kondo explains that in *Manyoshu*, which is thought to have been compiled in the late 8th century in the Nara period, many poems feature 'mountain' and 'the sea.' The AI analysis has drawn the firm conclusion that the trends in *waka* poetry underwent major transformations during the 700-year span between the late 8th and the 13th century: from 'mountains' and 'the sea' to 'birds' and 'flowers' to 'water' and 'night.'

By the way, a similar method can be applied to the contemporary Internet culture. "Take the Internet meme Kabosu-chan, a Shiba Inu (a breed of hunting dog from Japan), as an example. How did Kabosu-chan spread on X? What communities or user clusters on X were responsible for the meme going viral? How did it become an icon for Dogecoin (cryptocurrency)? I think it's possible to use AI vector analysis to answer these questions." Let's get back on track. AI analysis of the corpus may well shed new light on classical literature. The shift in the trends from 'mountains' and 'the sea' to 'birds' and 'flowers' to 'water' and 'night' might indicate some change in the ancient people's consciousness and language usage. The new picture of the Nara and Heian period that emerges through AI analysis may differ entirely from what we were taught in classical literature or history classes.

The most distinctive feature of *Nikkoku* is its example-based principal; the dictionary is compiled based on examples of actual use of words. This allows users to trace each word comprehensively from its first appearance to its historical changes. Even archaic words that are no longer used in modern language are brought to life through carefully selected examples that reveal their meanings and usages. *Nikkoku* is not only a language dictionary, but also a historical dictionary that illustrates the history of the Japanese language from the Nara period to the present day. The image of classical literature illuminated by the AI analysis of corpus linguistics may be reflected in the Third Edition's definitions and usage examples. This, in turn, may bring new value to its role as a historical dictionary. "That would be wonderful and I presume that's possible. With its planned release in 2032, it is our hope that the dictionary's editorial process will benefit from the latest advances in linguistic research."

What will *Nikkoku* be like in 2032? AI is no longer a tool reserved for dictionary editors and lexicographers alone; it is becoming an integral part of our society. ChatGPT tells us to some extent the meanings of words. And eight years from now, the precision and range of AI technologies will certainly far surpass what we know today. In such a future, what role should the dictionary play? That is the final question we pose to Kondo, as the interview comes to a close. "Digital texts have been increasing at a

tremendous rate. That looks like an explosion of the corpus. The Third Edition of Nikkoku will be crucial as a gateway into that vast corpus. It is like a 'basic ledger' of language. Each word is where it should be within the system of Japanese—a kind of proof of existence. More simply put, if a word appears in the dictionary, it is regarded as a legitimate word. This is *Nikkoku*'s greatest role, and it is something I believe AI cannot easily replace."

Translated by Matsui Mio. This article first appeared in the February 2025 no. 489 issue of Tokyojin *pp. 38–41, "AI to Amu Jisho: Kopasu ga Kotoba no Mirai ni Kakumei wo Okosu!? (Weaving a Dictionary with AI)." (Courtesy of Toshi Shuppan)* [April 2025]

KONDO Yasuhiro, Ph.D.

Professor Emeritus, Aoyama Gakuin University

Born in Gifu Prefecture in 1955. Graduated from the Department of Japanese Language and Literature, Graduate School of Humanities, The University of Tokyo. After working as an assistant professor at the Faculty of Letters, University of Tokyo and an associate professor at Japan Women's University, he became a professor at the Faculty of Letters, Aoyama Gakuin University. His specialties include grammar and corpus linguistics. He served as president of the Society for Japanese Linguistics from 2021 to 2024. His many publications include *Nihongo Kijutsu Bunpo no Riron* (Theory of Japanese descriptive grammar) and *Kopasu to Jisho* (Corpus and dictionaries) (co-authored).

MATSUI Mio

Editor and writer

Obtained an MSc in Digital Journalism from Goldsmiths, University of London. Worked as a manga editor at Kodansha for 10 years. Her editing work on *Kimi wa Pet* (Tramps Like Us) won the 27th Kodansha Manga Award. As a freelance editor/writer she has contributed to various magazines, including *Tokyojin*, *Da Vinci* and *With Online*.
