EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Building a Controlled Lexicon for Authoring Automotive Technical Documents

Miyata R., Sugino H.

Nagoya University, Japan

Abstract

We describe the framework and the process of building a controlled lexicon, specifically intended for authoring Japanese automobile repair manuals. Focusing on verbs, we seek to control two types of linguistic variations: (1) synonymous words and (2) case (argument) order variations. For synonymous words, we comprehensively extracted verb tokens from a large text data set and classified each verb type as approved or unapproved. For case order variations, we descriptively analysed case structures of Japanese sentences in the data set and defined the canonical order. We also examined the status of the constructed lexicon in terms of coverage, which enables us to establish a tangible goal of future lexicon building. The resultant controlled lexicon with 910 verbs and 954 case patterns can help authors choose appropriate words and construct consistent sentence structures. In order to accomplish effective and efficient authoring, we further proposed and designed two types of authoring support tools: a sentence diagnostic tool that identifies unapproved variations of verbs and sentence structures, and a template-driven writing tool that helps writers compose controlled sentences by completing canonical case patterns.

Keywords: controlled lexicon building; technical authoring; descriptive analysis; variation management; grammatical case; automotive domain

1 Introduction

Controlled lexicon, or controlled vocabulary, is a list of approved words in a certain domain, which may further determine unapproved words and provide the definition and usage of the registered words (Nyberg et al. 2003; Warburton 2014). The deployment of controlled lexicon helps enhance the consistent use of words—in particular verbs, nouns, adjectives and adverbs—in writing technical documents and prevents ambiguous and difficult expressions, which will lead to not only improved readability but also translatability of the documents. Furthermore, in combination with well-managed bilingual dictionaries, we can envisage the improved quality of machine translation outputs.

In this study, we describe the framework and the process of building a controlled lexicon specifically intended for authoring the Japanese automobile repair manuals of Toyota Motor Corporation. For every new model of automobile, huge volumes of technical documents, such as repair manuals, are created and translated, and an assemblage of writers and translators are involved in the document production workflow. This makes it difficult to ensure linguistic consistency across documents, eventually inducing a lack of clarity in the readers' understanding. In this context, we are now developing a controlled language for Japanese automotive technical documents. Controlled languages for authoring and translation basically consist of a syntactic and a lexical component (Nyberg et al. 2003; Kuhn 2014). In this paper, we report on the compilation of a controlled lexicon with specific focus on verbs, since verbs are crucial building blocks for composing operational instructions for repair manuals and governing sentence structures such as predicate-argument structures.

Although many controlled languages have been developed for particular domains, including the automotive domain (Means & Godden 1996; Godden 2000), few are publicly available. One of the few exceptions is ASD Simplified Technical English, or ASD-STE (ASD 2017), which was originally developed for aerospace maintenance documentation, and is now widely used in other industries. It defines writing rules that restrict certain syntactic/textual features, including sentence length and compound nouns, and provides a lexicon of approved and unapproved words. While ASD-STE is useful in its own right, it is not easy to directly port it to other purposes, domains and languages. In the case of controlled authoring of Japanese automotive technical documents, we also need extensive information of word usage. Thus, referring to the ASD-STE as a model example, we decided to build our controlled lexicon from scratch.

However, the practical problem is that few studies have established the general process of controlled lexicon building; in many cases, a controlled lexicon has been developed chiefly based on the tacit knowledge of domain experts and researchers. In this study, our lexicon building proceeded as follows: we first collected verb occurrences from existing texts, and then defined approved verbs and their canonical usage by analysing their occurrences. One of the important contributions of this paper is to document the controlled lexicon building process in detail, which will be helpful for related endeavors in the future.

The remainder of the paper is structured as follows. In Section 2, we design our controlled lexicon that enhances consistent writing of technical manuals. Section 3 describes the process of collecting verbs and controlling the variations to prepare a list of approved and unapproved verbs, presenting the growth of coverage in accordance with the building process. In Section 4, we further extend our controlled lexicon by defining the canonical case (argument) structure patterns for approved verbs. We then propose and design authoring support tools in Section 5 and conclude this paper with future outlook in Section 6.

2 Design of Controlled Lexicon

We address the two problems of inconsistent use of verbs: (1) synonyms and (2) case (argument) order. In automobile repair manuals, different verbs are sometimes used for the same operations, such as *koukansuru* and *torikaeru*, both of which mean 'replace'. These variations violate the basic principle of controlled lexicon, that is, 'one word – one meaning' (Nyberg et al. 2003; Møller & Christoffersen 2006), and may hinder readers' comprehension of the documents. Further, the notion of *case* (Fillmore 1968) is significant for controlled writing as Japanese case order is fairly free (Masuoka & Takubo 1992; Sasano & Okumura 2016), which sometimes causes structural variations. The following two sentences present an example of the different case orders of the Japanese verb *setsuzokusuru* (connect):

(1) GTS□ DLC3 □□□□□□/ GTS *o* DLC3 *ni* *setsuzokusuru*.
(2) DLC3 □ GTS □ □□□□□/ DLC3 *ni* GTS *o* *setsuzokusuru*.

The order of the accusative case (-*o*) and the dative case (-*ni*) is different from each other. Both sentences are grammatically correct in Japanese and can be translated as 'Connect GTS to DLC3'. They do not even necessarily hinder readers' comprehension of the text. From the viewpoint of consistent authoring; however, these variations should be avoided. In addition, if we can fully reduce these variations in the source, we can expect the improved results in parsing, text retrieval and machine translation.

Here, we propose a controlled lexicon that can enhance the consistency of writing in Japanese by extending the existing framework of controlled languages. To address the problem of synonyms, based on the ASD-STE, we create a list of approved and unapproved words with word definitions and examples. To address the problem of case order, we further define the canonical case order for each verb.

Figure 1 shows examples of approved and unapproved words in our controlled lexicon. Each entry word has the part of speech, semantic category and example sentence(s) of the word. Unapproved words have the links to the approved words, while approved words have definitions. These descriptions of the words help writers consistently select an appropriate word in writing text. Furthermore, the entries of approved words accompany the canonical case order(s) to support writers to appropriately construct sentences in a controlled manner.

Approved word	交換する/*koukansuru*
Part of speech	verb
Semantic category	Action > Part
Definition	'To remove an item and to install a new or serviceable item of the same type'. (ASD 2017)
Canonical case order	[PART/ITEM ***o***] [PART/ITEM ***ni***] *koukansuru*
Example	センサーを新品に交換する。/*Sensa o shinpin ni koukansuru.* (Replace the sensor with a new one.)
Unapproved word	取り替える/*torikaeru*
Part of speech	verb
Semantic category	Action > Part
Approved alternative	交換する/*koukansuru*
Unapproved example	必ず新品に取り替える。/*Kanarazu shinpin ni torikaeru.* (Always replace it with a new one.)

Figure 1: Examples of entries in the controlled lexicon: *koukansuru* and *torikaeru* (replace). For explanation, the definition of 'replace' is extracted from the specification of ASD-STE as a definition for *koukansuru*.

3 Construction of Controlled Lexicon

In this section, we elucidate how we collected approved and unapproved verbs from the text data of automotive technical documents. We also present the semantic categories of collected verbs and detailed analyses of the frequency of verb occurrence in the text data, which enables us to understand the status of lexicon in terms of coverage.

3.1 Verb Collection and Variation Control

From 17 sets of repair manuals that cover 10 types of automobiles from Toyota Motor Corporation, we comprehensively extracted verb tokens used in the main clauses of sentences, using Japanese sentence analysis tools JUMAN++V2 (Morita et al. 2015; Tolmachev et al. 2018) and KNP (Kawahara & Kurohashi 2006). We assume that they cover a sufficient range of verbs in this domain as we extensively investigated huge volumes of text data containing more than one million sentences. Subsequently, we eliminated verbs which were wrongly identified as verbs by the tools and rare compound verbs that can be replaced by simpler verbs. For example, *sokutei-kaishisuru* is a compound verb which combines the two simple expressions *sokutei* (measurement) and *kaishisuru* (start), and can be rephrased into *sokutei o* *kaishisuru* (start taking a measurement) or, simply, *sokuteisuru* (measure). We finally collected 1,058,424 verb tokens and 910 verb types.

The next task was to classify whether each verb (type) is approved or unapproved. We conducted the following steps:

1. Gather semantically similar verbs
2. If a verb is mostly interchangeable with another verb in actual sentences, regard them as verb variations
3. Define one of the verb variations as approved and the rest unapproved
4. Link the unapproved word(s) to the approved verb

In Step 3, we used the frequency of the words in the data set as an important evidence for decision-making; the more frequently the word occurs in the data, the more likely that it is approved.

Through this process, we finally identified 822 approved and 88 unapproved verbs. Table 1 presents the basic statistics of the constructed controlled lexicon, showing that approximately 10% of verb types and 3% of verb tokens (i.e. verb occurrences in our data set) were labelled as variations. It suggests that even documents that were authored by technical writers contained a certain amount of inconsistent use of verbs and we can expect an improvement in document consistency by employing our constructed lexicon.

Table 2 shows the verb variation categories with their frequencies in data set and examples. The major category is the synonym, i.e. a word that conveys almost the same meaning. We also regarded the use of prefix to verbs as variations. The prefixes are productive and potentially create many verb types, which is not desirable from the viewpoint of controlled writing. The important point is that prefixed verbs can be decomposed into a simple combination of a verb and an adverb as follows:

(3) □ □ □ □ □ □ □ □ □ □ □ □ □ **再確認する** □ / *Daiagunoshisu-kodo o* ***sai-kakuninsuru***.
(**Re-check** the diagnosis code.)
(4) □ □ □ □ □ □ □ □ □ □ □ □ □ **再び 確認する** □ / *Daiagunoshisu-kodo o* ***futatabi kakuninsuru***.
(**Check** the diagnosis code **again**.)

In this case, *sai-kakuninsuru* (re-check) can be decomposed into two simple words, *futatabi* (again) and *kakuninsuru* (check). We prohibited the following six types of prefixes and identified unapproved verbs: *sai-* (re-, again), *kari-* (temporary), *go-* (false), *zen-* (all), *ryo-* (both) and *shi-* (trial). Compound verbs combine two similar verbs. In many cases, it is possible to convert them into simpler verbs. The last category is the notational variation, which is peculiar to Japanese language. For example, the Japanese verb *atatameru* (warm up) can be written in different forms of kanji such as □ □ □ and □ □ □ . The meaning of the two notations are almost similar and can be controlled.

	All	Approved		Unapproved	
	#	#	%	#	%
Type	910	822	90.33	88	9.67
Token	1,058,424	1,027,659	97.09	30,765	2.91

Table 1: Statistics of the constructed controlled lexicon of verbs.

Category	#	Example (unapproved verb: corresponding approved verb)
Synonym	49	*jisshisuru* (conduct): *okonau* (do), *jokyosuru* (eliminate): *torinozoku* (remove)
Use of prefix	33	*sai-kakuninsuru* (re-check): *kakuninsuru* (check), *kari-koteisuru* (temporarily-fix): *torinozoku* (fix)
Compound verb	3	*ooi-kakusu* (cover and conceal): *oou* (cover), *arainagasu* (wash away): *arau* (wash)
Notation	3	□ □ □ : □ □ □ / *atatameru* (warm up), □ □ : □ □ □ / *wakaru* (know)

Table 2: Verb variation types with examples.

3.2 Semantic Category of Verbs

In conjunction with controlling verb variations, we analysed each verb and labelled a semantic category. Based on the lexicographic categories defined in WordNet,[1] which is often called *supersenses* (Ciaramita & Altun 2006; Paaß & Reichartz 2009), we created a typology of semantic categories of verbs. Table 3 presents the typology of a semantic category of verbs with examples and frequency information. While WordNet defines 15 categories for verbs, namely, body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative and weather, we adopted the necessary classification for the automotive technical documentation and modified them as necessary to create top level categories. Second level categories were defined in a bottom up manner through the analysis of verbs.

All the 910 verb types were classified into one of the semantic categories. The dominant category is the Action > Part. The verbs in this category are used to express operational actions regarding automobile parts, which are the core building blocks of procedural instructions for automobile repair tasks. These semantic categories can help writers choose appropriate verbs when authoring documents.

[1] https://wordnet.princeton.edu

Level 1	Level 2	Example	Type		Token	
			#	%	#	%
Action	Part	*toritsukeru* (install), *kirihanasu* (disconnect)	343	37.7	395,890	37.40
	Software	*hyojisuru* (display), *kiokusuru* (store)	54	5.9	97,944	9.25
	Diagnosis	*tenkensuru* (check), *kanshisuru* (monitor)	23	2.5	96,457	9.11
	General	*shiyousuru* (use), *sousasuru* (operate)	91	10.0	50,436	4.77
	Auxiliary	*suru* (do, make), *okonau* (perform)	10	1.1	189,140	17.87
Stative	Existence	*aru/iru* (be, exist), *ichisuru* (be located)	11	1.2	41,544	3.93
	Composition	*kouseisuru* (compose), *yuusuru* (have)	21	2.3	10,855	1.03
	Denotation	*shimesu* (indicate), *arawasu* (show, denote)	18	2.0	8,606	0.81
	Relation	*kankeisuru* (be related), *kotonaru* (differ)	28	3.1	2,439	0.23
	Function	*kinousuru* (function), *eikyosuru* (affect)	31	3.4	5,084	0.48
	State	*taikisuru* (wait), *nokoru* (remain)	25	2.7	18,931	1.79
	Auxiliary	*hajimeru* (start), *keizokusuru* (continue)	29	3.2	8,195	0.77
Change	State	*modosu* (return), *hasseisuru* (occur, generate)	65	7.1	7,305	0.69
	Quantity	*atatameru* (warm up), *joshosuru* (increase, rise)	34	3.7	3,967	0.37
	General	*henkasuru* (change), *hendousuru* (vary)	5	0.5	4,326	0.41
Communication	Exchange	*soushinsuru* (send), *tsuuchisuru* (inform)	36	4.0	6,869	0.65
	Record	*kirokusuru* (record), *hozonsuru* (save, store)	14	1.5	5,764	0.54
	Performance	*kinshisuru* (prohibit), *shitagau* (follow)	22	2.4	2,717	0.26
Cognition		*chuuisuru* (pay attention to), *handansuru* (judge)	44	4.8	101,541	9.59
Perception		*miru* (see), *kanjiru* (feel), *kiku* (hear)	6	0.7	414	0.04

Table 3: Typology of the semantic categories of verbs with examples and frequency information.

Rank	Verb	Frequency		Cumulative Frequency (Coverage)	
		#	%	#	%
1	*suru* (do, make)	94,613	8.94	94,613	8.94
2	*okonau* (perform)	93,750	8.86	188,363	17.80
3	*kakuninsuru* (confirm)	87,099	8.23	275,462	26.03
4	*torihazusu* (remove)	70,011	6.61	345,473	32.64
5	*kirihanasu* (disconnect)	65,451	6.18	410,924	38.82
6	*toritsukeru* (install)	64,508	6.09	475,432	44.92
7	*setsuzokusuru* (connect)	55,546	5.25	530,978	50.17
8	*sokuteisuru* (measure)	50,703	4.79	581,681	54.96
9	*tenkensuru* (check)	44,254	4.18	625,935	59.14
10	*aru* (be)	40,728	3.85	666,663	62.99
11	*koukansuru* (replace)	23,614	2.23	690,277	65.22
12	*sentakusuru* (select)	19,632	1.85	709,909	67.07
13	*taikisuru* (wait)	18,448	1.74	728,357	68.82
14	*shoukyo* (clear)	16,736	1.58	745,093	70.40
15	*shutsuryokusuru* (output)	14,171	1.34	759,264	71.74
16	*hyojisuru* (be displayed)	10,499	0.99	769,763	72.73
17	*yomu* (read)	10,004	0.95	779,767	73.67
18	*shiyousuru* (use)	8,989	0.85	788,756	74.52
19	*sadousuru* (operate)	7,301	0.69	796,057	75.21
20	*naru* (become)	7,003	0.66	803,060	75.87

Table 4: The 20 most frequent controlled verbs that occurred in our data set.

3.3 Coverage of Verbs

Currently, the number of approved verb types in our lexicon is 822. However, from the practical point of view, it is still too many and writers—even professional technical writers—may find it difficult to appropriately make use of the controlled lexicon. Here, it is more valuable to define the core set of verbs, or further reduce the number of verb entries. To set the goal of the lexicon refinement process, in this section, we investigate the coverage of verbs and estimate how many verbs are necessary for authoring automotive technical documents.

Table 4 presents the 20 most frequent controlled verbs that occurred in our data set with their individual frequencies and cumulative frequencies, which can be regarded as coverage. It is worth noting that only seven verbs cover half of the verb use in the automotive repair manuals and 20 verbs three quarters. Figure 2 illustrates the growth curve of coverage as the number of verbs increases in the descending order of frequency. The curve grows rapidly by around 100, then tapers off and, around 400, is flattened out. From this observation, we can recognize that approximately 100 verb types are particularly important and can be considered a core set of verbs, and 300–400 verb types will suffice for our controlled lexicon. Further refinement of the lexicon will be part of future research.

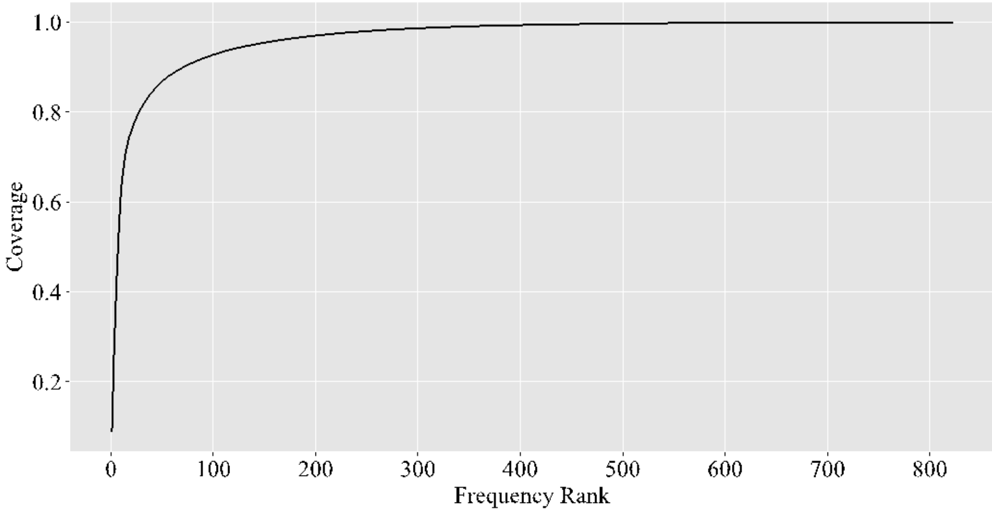

Figure 2. Growth curve of the coverage of the controlled verb lexicon.

4 Formulation of Canonical Case Order

4.1 Procedure

To further increase the utility of our controlled lexicon, we formulated the canonical case orders for all the approved verbs in the following two processes.

1. For each verb, we abstracted the case structure of sentences (main clauses) using Japanese sentence parser KNP with JUMAN++V2 (see also Figure 3). For each sentence, a verb at the end of the sentence, which is a predicate of the main clause in Japanese, was identified. Subsequently, noun phrases with postpositional particles, such as *ga*, *o*, *ni* and *de,* that directly attach to the verb were extracted as cases, or arguments, for the verb. Only the particles were reserved to form an abstract case pattern. At this stage, supplementary adverbial phrases were omitted because they were usually irrelevant to the core structure of sentences.
2. We selected the preferred case orders based on the frequency of usage and combined them to create canonical orders that could cover frequent types of the case structures for the verb (see also Figure 4). It should be noted that there were many less frequent types that are not covered by the defined case orders, which we will discuss in the next section.

After conducting these processes for all the 822 approved verbs, we finally formulated 954 canonical case orders. Some of the verbs have multiple canonical case structures; for example, the verb *hyojisuru* (display) has two structures: [~*o*] [~*ni*] *hyoji-saseru* and [~*ni*] [~*ga*] *hyoji-sareru*. Here are some examples of such sentences:

(5) □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ **表示さ せる** □ / *Daiagunoshisu-kodo kakunin gamen o* ***hyoji-saseru***.
(**Display** the diagnosis code check screen.)

(6) □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ **表示さ れる** □ / *Gamen ni daiagunoshisu-kodo ga* ***hyoji-sareru***.
(The diagnosis code **will be displayed** on screen.)

The verb *hyojisuru* can be used both in causative form *hyoji-saseru* for human actions and in passive form *hyoji-sareru* for machinery functions, and these two forms have different case structures. The detailed description of the usage of the different verb forms and case structures will be an important future task.

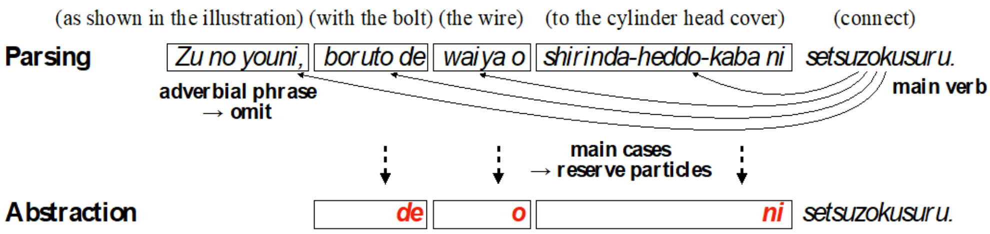

Figure 3. Automatic process to abstract the case structure of a Japanese sentence.

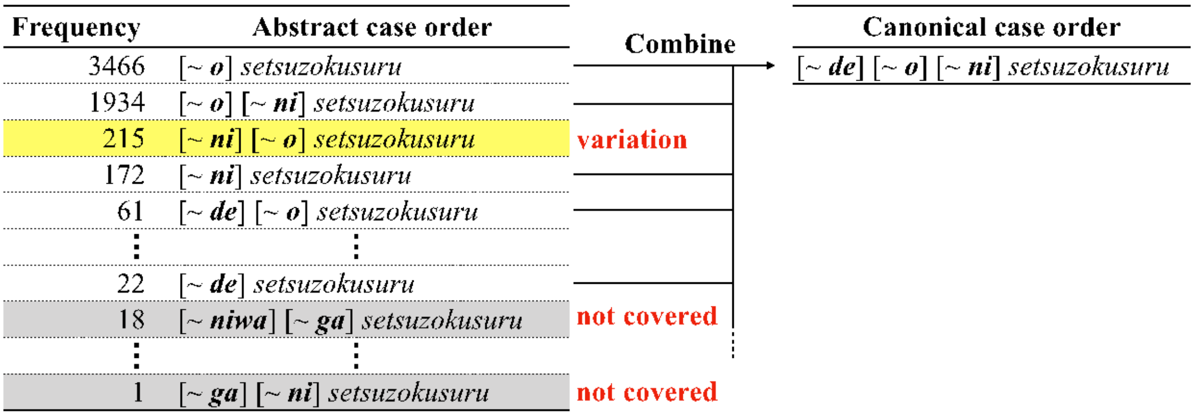

Figure 4. Examples of the formulation of canonical case order: *setsuzokusuru* (connect).

4.2 Coverage of Defined Case Structure

As mentioned above, it is difficult to comprehensibly cover all the case order patterns, although we assume that the formulated canonical case structures covered a substantial portion of case patterns. Here, we calculate the coverage of the 954 canonical patterns using the same data set of the automobile repair manuals. For each sentence, we abstracted the case pattern in the same manner described in Figure 3. If the abstracted case pattern contained the same set or a subset of case elements defined in the canonical structure, we regarded it as the covered pattern. In addition, if the order of the case elements violates the canonical order, we considered it as a variation.

Table 5 shows the results; 85.61% of the pattern tokens were covered by our formulated patterns, which demonstrated the fairly high coverage. However, the coverage of case pattern types is low. It indicates that a large number of rare types of case patterns have not been captured by the current set of patterns. The coverage needs to be increased by defining other types of canonical orders.

To understand the relationship between the number of case structure types and coverage of tokens, we observe the growth curve in Figure 5. This figure illustrates how coverage increases as case pattern types are included in the descending order of frequency. We can see that the 2,000 most frequent types can cover almost all tokens in our data set. The formulated canonical case patterns already covered 1,807 types, while they do not necessarily include the frequent ones. We assume we can soon attain higher coverage, namely more than 90%, by adding frequent patterns.

	All	Covered		Variation	
	#	#	%	#	%
Type	5,363	1,807	33.69	170	9.41
Token	1,058,424	906,134	85.61	24,366	2.69

Table 5: Coverage of the set of canonical case structures and ratio of variation in our data set.

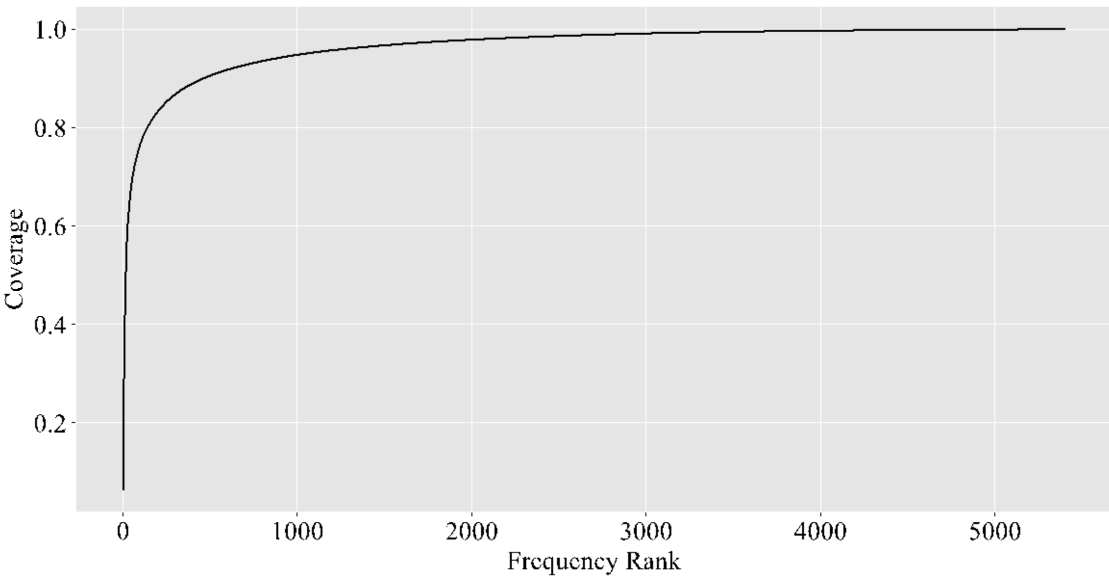

Figure 5. Growth curve of coverage of case structure patterns in the original forms.

4.3 Case Structure Variation

Finally, we investigate the issue of structural variations in our data set. As Table 5 shows, 2.7% of the covered pattern tokens were found to be variations. Although the ratio is not high, given that the automobile manuals were reasonably controlled by professional writers, there is still room for improvement in terms of consistent use of sentence structures. An example of case structure variations for the verb *tofusuru* (apply, coat) is shown below:

(7) □□□O□□□□ □□□□□□□□**塗布する**□ / *Shinpin no O-ring ni conpuressa oiru o* ***tofusuru***.
(**Apply** compressor oil to a new O-ring.)
(8) □□□□□□ SST □□□□□□ **塗布する**□ / *Gurisu o SST no boruto ni* ***tofusuru***.
(**Apply** grease to the SST bolts.)

Example (7) conforms to the defined canonical pattern, [~ *de*] [~ *ni*] [~ *o*] *tofusuru,* and Example (8) is regarded as a variation since the order of the [~ *o*] case and [~ *ni*] case is the reverse of the canonical order. In this example, we can modify (8) into (9) as shown below without changing its meaning.

(9) SST □□□□□ □□□□□□**塗布する**□ / SST no boruto ni gurisu o **tofusuru**.
(**Apply** grease to the SST bolts.)

Although (8) is grammatically correct and the core meaning of the sentence remains unchanged, it is important to control these structural variations for consistency, which will eventually lead to high usability of the text. Importantly, if we define canonical patterns in advance, these variations can be automatically detected in combination with sentence analysis tools, which we will discuss in the next section.

5 Towards Authoring Support

The constructed controlled lexicon of verbs with the definitions of canonical case orders is a basis for controlled authoring; it helps writers recognize which verb should be used in what sentence structure. To be more effective, we will further discuss the mechanisms for supporting the controlled authoring process of writers. Authoring support scenarios can be broadly divided into two types: *post hoc revision* and *writing from scratch*. Correspondingly, we propose a sentence diagnostic tool for revision and a template-driven writing tool based on the canonical case patterns. In the following two sections, we outline them respectively.

5.1 Diagnostic Tool

The constructed lexicon can be used to control the two types of variations, that is, the use of unapproved words and non-canonical sentence structures. We propose a tool to support the diagnosis and revision of both types of variations in three processes: *detect*, *suggest* and *rewrite*.
With regard to unapproved words, the three processes can be simply implemented if the synsets of the unapproved and approved words are defined (Warburton 2014). The tool first searches the input sentence for unapproved words referring to the lexicon and, if any unapproved word is discovered, it retrieves the corresponding approved word. If the suggestion is adopted, the unapproved word in the input sentence is automatically corrected.
Figure 6 depicts the process of detection and suggestion of unapproved words using our controlled lexicon checker. The following example is used as an input:

(10) □□□□□□□□□□□ **除去する**□ / *Shinpin no kawaita nuno nado de ibutsu o* ***jyokyosuru***.
(**Eliminate** foreign matter with a new dry cloth.)

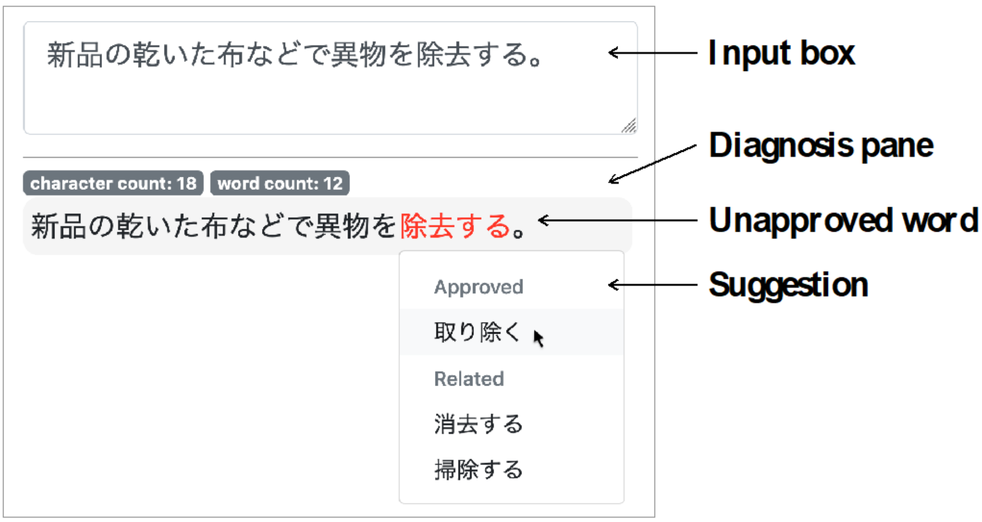

Figure 6. Prototype interface of the controlled lexicon checker.

The unapproved word *jokyosuru* (eliminate) is highlighted in red and the approved word *torinozoku* (remove) is recommended. To further support human decision-making, this tool also provides semantically-related words *shoukyosuru* (delete) and *soujisuru* (clean), which might be more suitable in a certain context. Writers can select any of the candidates to replace the unapproved word.
Conversely, case structure variations are more difficult to handle automatically. The first bottleneck is the parsing of the input sentence to abstract its case structure. Although the high-performance Japanese parsers, such as KNP (Kawahara & Kurohashi 2006) and CaboCha (Kudo & Matsumoto 2002), are available, the accuracy of parsing of long, complex sentences is still not sufficient for this task. Another difficulty is the ambiguity of suggestion. The canonical case patterns we defined do not specify the order of supplementary elements such as adverbial phrases. Even if a sentence is correctly parsed and the violated case order is detected, there may be multiple suggestions for rewriting. While the final decision making will be left to human writers, we need to further elaborate rules to place elements of sentence in proper positions.

5.2 Template-driven Writing Tool

As a more preemptive solution for possible violations to the controlled lexicon, we also plan to develop a template-driven writing tool. Here, we explain the design principle of the tool and challenges for its development.
The basic flow of template-driven writing is shown in Figure 7. To write instructional sentences, verbs are the most important elements that govern the core meanings and sentence structures. Thus, writers first select a main verb from the controlled lexicon of verbs. At this stage, the tool helps them to select an appropriate verb by displaying the hierarchic structure of semantic categories of verbs defined in Section 3.2. Once the verb is fixed, registered sentence templates, i.e. canonical case patterns for the verb, will be presented and writers can choose one of them. However, the default sentence template may not be sufficient for accommodating necessary information. Hence, writers can further add supplementary sentence components with slots, such as an adverbial component [~ *youni*] (as ~), from a list of possible components provided by the tool. In conjunction with populating the template with additional components, writers fill in the slots with content words such as *boruto* (the bolt), *waiya* (the wire) and *shirinda-heddo-kaba* (the cylinder head cover) to complete the sentence. At this stage, the tool provides two functions: (1) suggestion of the probable candidate words and technical terms and (2) validation of the conformity of content words to the controlled lexicon of nouns and adjectives.

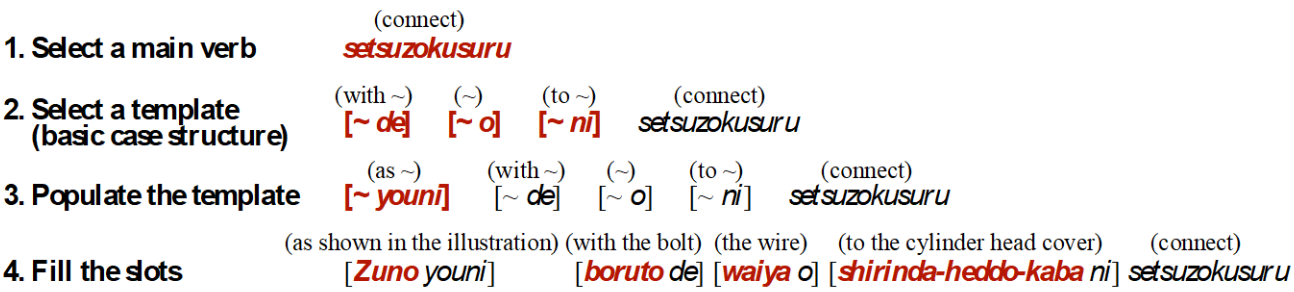

Figure 7. Basic flow of template-driven writing with a simple sentence as an example.

To implement the functions described above, the following items are specifically necessary:

1. Definition of supplementary sentence components (specifically, adverbial phrases), besides basic case patterns.
2. Construction of controlled lexicon of content words (specifically, nouns and adjectives), besides verbs.
3. Construction of controlled terminology of automotive domain (specifically, part and tool names).

Furthermore, to fully implement the template-driven writing tool, we need to tackle the challenges of constructing compound/complex sentences by combining multiple simple sentences. To understand how sentences are constructed in our data set, focusing on coordinate and adverbial clauses, we first automatically extracted compound/complex sentences and discovered that 525,966 of the 1,058,424 sentences are compound/complex. Sentence (11) is an example of a compound sentence, with two coordinate independent clauses, and Sentence (12) is an example of a complex sentence, with one main clause and one subordinate clause.

(11) エンジン□□□**し**□□□□□□□□□□□□□/
*Enjin o sidou-**shi**, aidoringu joutai ni suru.*
(Start the engine **and** keep the engine idling.)
(12) □□□□□□□□□□□□□**前に**□□□□□□□□□□□□□□□□□□□/
*Toraburu-shuto o jisshisuru **maeni**, kono kairo no hyuzu no tenken o suru-koto.*
(Inspect the fuses for circuits **before** performing the troubleshooting.)

As shown in bold in the examples above, certain connective expressions are used to construct compound/complex sentences. We categorized surface connectives automatically extracted from 525,966 compound/complex sentences. Table 6 shows the results of the categorization with frequency for each category. The total number of connectives is 739,651, which means that many of the compound/complex sentences have more than two clauses. The dominant category is the resultative coordination such as (11), occupying nearly 60%. We also notice that adverbial clauses to express timing and conditions of events (‘when’, ‘in case’ and ‘if’) appear frequently in the data set, which indicates that conditional branching of tasks are crucial building blocks for authoring instructional documents. Based on the results, we plan to implement functions to support the construction of compound/complex sentences.

Finally, from the viewpoint of controlled authoring, we should emphasize that various surface connectives are used to signify almost the same meaning. For example, to mean ‘when’ in adverbial clauses, various connectives are used, such as *toki*, *tokini*, *tokiwa*, *tokiniwa*, *sai*, *saini*, *saiwa* and *sainiwa*. In many cases, these connectives are interchangeable. Therefore, it is effective to define the approved usage of connectives to further control the sentence structural variations.

Level 1	Level 2	Level 3	Surface connectives	#	%
coordinate	resultative	and	V (continuative form), V-*te*	421,324	56.96
		such as	*tari*	6,389	0.86
	contradictory	but	*ga*	4,563	0.62
adverbial	time	when	*toki (ni/wa/niwa), sai (ni/wa/niwa)*	39,051	5.28
		each time	*tabi (ni)*	101	0.01
		before	*mae (ni)*	17,946	2.43
		after	*nochi (ni), ato (ni/de)*	12,305	1.66
		then	*ue (de)*	1,526	0.21
		until	*made*	10,805	1.46
	condition	in case	*baai (ni/ha/niwa)*	90,773	12.27
		if	*to, nara, ba*	37,373	5.05
		only if	*dakedemo*	210	0.03
		though	*mo*	6,491	0.88
		as long as	*kagiri*	102	0.01
	method	by	*kotode (mo/niyori)*	9,198	1.24
	attendant circumstances	while, with	*nagara, mama, tsutsu*	9,426	1.27
		without	*zu ni, nai de*	3,394	0.46
	state	in the state	*jotai de*	12,273	1.66
		in the way	*youni*	11,013	1.49
	purpose	in order that	*tame*	27,325	3.69
	reason	because	*tame (ni), node, kekka, kara*	17,779	2.4
	contradictory	but	*noni*	270	0.04
	extent	to the extent	*hodo*	14	0.00

Table 6: Compound/complex sentence patterns and surface connectives observed in our data set.

6 Conclusion

In this study, we have built the controlled lexicon of verbs that is useful for consistent authoring of automotive technical documents. The lexicon building proceeded in both descriptive and prescriptive manners: we first descriptively observed a huge volume of existing text data, and then prescriptively defined approved words and their canonical usage. Although we dealt with Japanese verbs as a starting point, this lexicon building process is applicable to other lexical units and languages.

Currently, the constructed lexicon consists of 822 approved and 88 unapproved verbs, which comprehensively cover the analysed data set containing more than one million verb tokens. The detailed analysis of coverage revealed that we can reduce the size of the lexicon to 300–400 words with little loss of coverage. The significant feature of our lexicon is the definition of canonical case orders for each approved verb, which helps writers compose sentences in consistent structures. We have defined 954 canonical case patterns that are estimated to cover 85% of the existing sentences.

We have also proposed authoring support tools that employ controlled lexicon. For two different scenarios, that is, post hoc revision and writing from scratch, we designed a sentence diagnostic tool and a template-driven writing tool, respectively. These tools are designed to assist writers in using appropriate words in accordance with their controlled usage. We also discussed necessary components and technologies to implement these tools.

In future research, we will refine the constructed lexicon and extend it to cover nouns, adjectives and adverbs. We also plan to build an English lexicon and link the approved words between Japanese and English, which enables consistent translation by both human translators and machine translation systems. In particular, we will examine the effectiveness of controlled bilingual lexicon for machine translation. We assume that the use of controlled lexicon can facilitate the reduction of vocabulary size of the text, which may have a positive impact on machine translation, including recent neural models. Finally, the implementation and evaluation of the authoring support tools we designed is an important practical goal of this research project, which we will address in future studies.

7 References

ASD. (2017). ASD Simplified Technical English. Specification ASD-STE100, Issue 7. http://www. asd-ste100.org [08/07/2020].

Ciaramita, M. & Altun, Y. (2006). Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 594–602.
Fillmore, C. J. (1968). The Case for Case. In Bach, E. & Harms, R. T. (eds.) *Universals of Linguistic Theory*, 1–88. New York: Holt, Rinehart and Winston.
Godden, K. (2000). The Evolution of CASL Controlled Authoring at General Motors. *Proceedings of the 3rd International Workshop on Controlled Language Applications*, Seattle, WA, USA, 14–19.
Kawahara, D. & Kurohashi, S. (2006). A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York, USA, 176–183.
Kudo, T. & Matsumoto, Y. (2002). Japanese Dependency Analysis using Cascaded Chunking. *Proceedings of the 6th Conference on Natural Language Learning*, Stroudsburg, Pennsylvania, USA, 63–69.
Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1): 121–170.
Masuoka, T. & Takubo, Y. (1992). *Kiso Nihongo bunpo* [*Basic Japanese Grammar*]. Tokyo: Kuroshio Shuppan.
Means, M. & Godden, K. (1996). The Controlled Automotive Service Language (CASL) Project. *Proceedings of the 1st International Workshop on Controlled Language Applications*, Belgium, Leuven, 106–114.
Morita, H., Kawahara, D. & Kurohashi, S. (2015). Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2292–2297.
Møller, M. H. & Christoffersen, E. (2006). Building a Controlled Language Lexicon for Danish. *LSP & Professional Communication*, 6(1): 26–37.
Nyberg, E., Mitamura. T. & Huijsen, W. O. (2003). Controlled Language for Authoring and Translation. In Somers, H. (ed.) *Computers and Translation: A Translator's Guide*, 245–281, Amsterdam: John Benjamins.
Paaß, G. & Reichartz, F. (2009). Exploiting Semantic Constraints for Estimating Supersenses with CRFs. *Proceedings of the SIAM International Conference on Data Mining*, Sparks, Nevada, USA, 485–496.
Sasano, R. & Okumura, M. (2016). A Corpus-Based Analysis of Canonical Word Order of Japanese Double Object Constructions. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2236–2244.
Tolmachev, A., Kawahara, D. & Kurohashi, S. (2018). Juman++: A Morphological Analysis Toolkit for Scriptio Continua. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, System Demonstrations*, Brussels, Belgium, 54–59.
Warburton, K. (2014). Developing Lexical Resources for Controlled Authoring Purposes. *Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use*, Reykjavik, Iceland, 90–103.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 19K20628 and 19H05660, and by the Naito Research Grant, Japan. The automobile manuals used in this study were provided by Toyota Motor Corporation.