https://doi.org/10.14618/ids-pub-11201

Takahiro Makino / Rei Miyata / Seo Sungwon / Satoshi Sato

DESIGNING AND BUILDING A JAPANESE CONTROLLED LANGUAGE FOR THE AUTOMOTIVE DOMAIN

Toward the development of a writing assistant tool

Abstract In this paper, we propose a controlled language for authoring technical documents and report the status of its development, while maintaining a specific focus on the Japanese automotive domain. To reduce writing variations, our controlled language not only defines approved and unapproved lexical elements but also prescribes their preferred location in a sentence. It consists of components of a) case frames, b) case elements, c) adverbial modifiers, d) sentence-ending functions, and e) connectives, which have been developed based on the thorough analyses of a large-scale text corpus of automobile repair manuals. We also present our prototype of a writing assistant tool that implements word substitution and reordering functions, incorporating the constructed controlled language.

Keywords Japanese controlled language; corpus-based lexicon building; variation management; writing support tool; automotive domain

1. Introduction

The production process of technical documents for industrial products, such as automobile repair manuals, usually involves many writers and editors. This induces writing variations, which might not only degrade the searchability of document content for readers, but also the reusability of past text for writers. These variations may also have a negative impact on translation memory tools, degrading the reusability of past translations.

To reduce such variations, it is crucial to use a properly designed controlled language in combination with a writing assistant tool. Controlled languages restrict the syntax and/or lexicon of a certain natural language (Kittredge 2003; Kuhn 2014). The syntactic restrictions are usually defined through writing rules, such as ‘write short and clear sentences’ (ASD 2021). Although such rules provide a general guide to writing, some may not concretely indicate how to compose a sentence. The lexical restrictions take the form of a controlled lexicon, which is a list of approved words. The lexicon becomes more useful if unapproved words are linked to approved words (Warburton 2014). Another challenge is the provision of detailed descriptions of word usage. For languages with flexible word order, such as Japanese, the regulation of word locations in sentences is important when improving textual consistency.

While many English controlled languages have been proposed and used in practice (Kuhn 2014), the development of Japanese controlled languages has not advanced sufficiently. Current approaches to Japanese controlled languages for writing purposes mainly address syntactic restrictions (e. g., Japan Technical Communicators Association 2016; Japio 2018). Furthermore, the tools for assisting Japanese controlled writing are scarce (Miyata et al. 2016).

Against this backdrop, domain-specific lexical restrictions are needed to properly manage the writing variations. Miyata/Sugino (2020) reported the building process of a Japanese controlled lexicon for the automotive domain, specifically focusing on the verbs and their

This paper is part of the publication: Klosa-Kückelhaus, Annette / Engelberg, Stefan / Möhrs, Christine / Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag. https://doi.org/10.14618/phpy-6r66

case orders. To further cope with various writing variations, we need to extend the scope to cover other lexical elements necessary for writing sentences. In this study, therefore, we design and build a Japanese controlled language for writing technical documents in the automotive domain that covers a wide range of linguistic elements. We also introduce our prototype tool designed to help writers reduce various types of writing variations.

In section 2, we propose our controlled language with its design principle, components, and the general methodology to build it. We then present the three of the components in sections 3–5, respectively, showing the detailed building process and results. In section 6, we introduce a prototype of our writing assistant tool that implements part of the controlled language. Finally, we conclude this paper with implications for future work in section 7.

2. Design of controlled language

2.1 Principle

To avoid writing variations, our controlled language should comply with the principle 'one meaning/function should correspond to one form'. Some controlled languages, including ASD-STE 100 (ASD 2021), are designed to comply with the reverse ('one form should correspond to one meaning/function') as well because polysemy might lead to the ambiguity of reading and hinder the readers' understanding. Although we did not count this as a requirement, we aimed to achieve this when building a controlled language.

As mentioned in section 1, for the purpose of controlled writing, it is effective to include unapproved words in the lexicon and link them to approved words (Warburton 2014, 2021). ASD-STE 100, for example, regulates the use of an unapproved verb 'delete' and provides its alternatives, namely, 'disconnect', 'disengage', and 'remove', each of which is defined to have a unique approved meaning. The comprehensive prescriptions of such linkages between unapproved and approved words are possible if the target domain is sufficiently specified. Like ASD-STE 100, which originally focused on aerospace maintenance documentation, our controlled language is intended for a specific domain. Thus, we reasonably assume that we can widely define the unapproved words in addition to approved ones.

Another important principle of our controlled language is that syntactic restrictions are incorporated in the lexical components. Previous controlled lexica often do not explicitly specify the detailed syntactic information. ASD-STE 100, for example, provides an approved word 'remove' with its part-of-speech information ('v'), approved meaning ('To take or move something away from its initial position'), and approved example ('REMOVE THE INDICATOR FROM THE PANEL.') (ASD 2021, p. 2-1-R9). However, for languages with flexible word order, more detailed information might be needed to avoid writing variations. In Japanese, for example, the following two sentences are grammatically correct:

(1) インジケータを パネルから 取りはずす。/ *Injiketa o paneru kara torihazusu.*

(2) パネルから インジケータを 取りはずす。/ *Paneru kara injiketa o torihazusu.*
(Translation: Remove the indicator from the panel.)

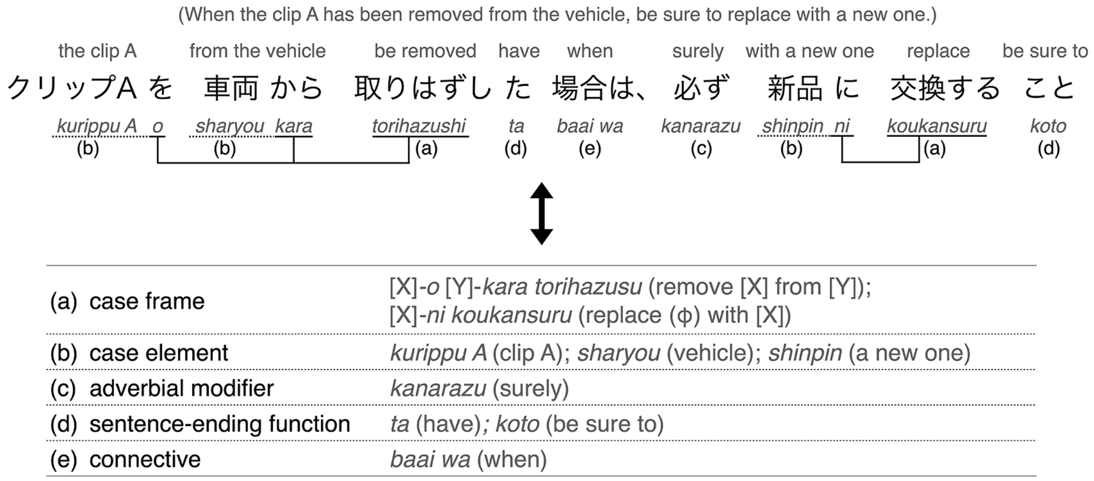
In this example, the case order for the verb *torihazusu* (remove) is swapped between (1) and (2), without changing the sentence's meaning. To control such variations, it would be effective to provide preferred word order information in the entries of the lexicon.

2.2 Components

Although Miyata/Sugino (2020) already compiled a controlled lexicon of verbs, they did not cover other parts of speech, such as nouns and adverbs. Therefore, to control the extensive range of writing variations, we designed a controlled language that consists of the following five components:

(a) **case frames:** verbs with canonical orders for their argument slots (Miyata/Sugino 2020)
(b) **case elements:** nouns or noun phrases that can fill argument slots
(c) **adverbial modifiers:** adverbs or adverbial phrases with their preferred locations
(d) **sentence-ending functions:** sequences of functional words attached to sentence-ending main verbs
(e) **connectives:** conjunctions and conjunctive phrases that indicate inter-clause or inter-sentence relationships

Figure 1 shows an example of a Japanese sentence annotated using our controlled language. Most sentences in our target documents can be broken down into elements derived using the five components.



(a) case frame	[X]-*o* [Y]-*kara torihazusu* (remove [X] from [Y]); [X]-*ni koukansuru* (replace (ϕ) with [X])
(b) case element	*kurippu A* (clip A); *sharyou* (vehicle); *shinpin* (a new one)
(c) adverbial modifier	*kanarazu* (surely)
(d) sentence-ending function	*ta* (have)*; koto* (be sure to)
(e) connective	*baai wa* (when)

Fig. 1: Sentence annotated with five types of elements defined in our controlled language

It is significant that each entry in these components is not only a single word but also a sequence of words. For example, the connective *baai wa* (when) in Figure 1 is composed of a noun *baai* and a particle *wa*, and the combination of the two words can be regarded as a basic operational unit when writing sentences. While the existing controlled lexica tend to register single words as entries, we flexibly define the linguistic spans of lexicon entries with the purpose of controlled writing assistance.

Here, we specifically provide the details of components (c)–(e) in sections 3–5, respectively, which have not been sufficiently investigated in previous studies.[1]

[1] The component (a) was compiled by Miyata/Sugino (2020). The component (b) virtually means the terminology of the target domain. The terminology management for controlled authoring has been discussed by Warburton (2021).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

2.3 Methodology for controlled language building

We adopted a corpus-based method to build a controlled language. The corpus was first constructed by extracting sentences from 17 sets of automobile repair manuals provided by Toyota Motor Corporation. A total of 1,053,111 sentence tokens (158,383 sentence types) were collected (henceforth, **Corpus-Org**). As the Corpus-Org includes complex/compound sentences with more than one clause, we then decomposed them into simpler sentences using a Japanese sentence splitting tool (Kato et al. 2020a) and obtained 1,428,381 sentence tokens (159,816 sentence types) (henceforth, **Corpus-Simple**).

We built each component using the following steps: i) comprehensively collecting instances from the corpus; ii) grouping the instances in terms of meaning/function; iii) defining approved and unapproved types for each group; and iv) specifying the preferred usage of the approved types, if any.

In step ii), we specifically examined the *interchangeability* of instances in actual sentences in the corpus. The following examples shows the interchangeability of adverbial modifiers.

(3) 空気が極力入らないように張り付ける / *Kuuki ga kyokuryoku hairanai youni haritsukeru.*

(3’) 空気がなるべく入らないように張り付ける / *Kuuki ga narubeku hairanai youni haritsukeru.*

(Translation: Attach so that air does not enter as much as possible.)

Here we observe that the word *kyokuryoku* (as much as possible) can be replaced with the word *narubeku* without changing the meaning of sentence (3). It should be noted that synonymous words obtained from general thesauri do not guarantee their interchangeability in the target domain text. In addition, words that are judged as interchangeable may not be regarded as synonymous in a strict sense.

In steps iii) and iv), to define the approved types or preferred usage, we chiefly referred to frequency information; the more frequently observed in the corpus, the more likely to be specified as approved/preferred. Since the previously-authored documents are primarily referred to by writers as ‘models’ in the writing process, it is reasonable to use the frequency information in the target corpus as significant evidence for defining approved words and preferred usage.

3. Lexicon of adverbial modifiers

The adverbial modifiers were divided into two types: fixed and variable. The former includes adverbs, such as 必ず/*kanarazu* (surely), and adverbial phrases, such as 同量ずつ/*douryouzutsu* (by an equal amount). The latter includes adverbial phrases with a slot, such as 約[X]分間/*yaku* [X] *funkan* (for approximately [X] minutes).

3.1 Building procedure

To widely collect instances of adverbial modifiers, we first manually investigated the 200 most frequent sentences in the Corpus-Org and defined heuristic rules to extract instances based on the parsed results of Japanese sentence analysis tools, Juman++ (Tolmachev/Kawahara/Kurohashi 2018) and KNP (Kawahara/Kurohashi 2006). Table 1 shows the formulated linguistic rules; they are defined based on the types of last morpheme (if it is a particle, the second last) of each modifier of the main verb. Using these rules, we automatically collected 3,922 types of adverbial modifier candidates.

No	Last morpheme type	Extracted example
1	adverb	ゆっくり/*yukkuri* (slowly)
2	continuative form of adjective	無理に/*murini* (by force)
3	temporal noun that can be used as an adverb	再度/*saido* (again)
4	suffix to make an adverb: 的に/*tekini*, めに/*meni*	定期的に/*teiki-tekini* (periodically)
5	temporal suffix: 後/*go* (after), 中/*chu* (while), 前/*mae* (before), 時/*ji* (when)	作業前に/*sagyou mae-ni* (before proceeding with work)
6	numeral suffix: 回/*kai* (time), 秒間/*byoukan* (second) ,秒/*byou* (second), 分間/*funkan* (minute), 分/*fun* (minute), %, 種類/*syurui* (kind), 以上/*ijou* (more than), ずつ/*zutsu* (one by one), 程度/*teido* (about), 程/*hodo* (about)	15秒間/*15 byoukan* (for 15 seconds)

Table 1: Linguistic rules to extract adverbial modifiers

We then manually classified them into fixed and variable types. The instances extracted by rules 5 and 6 in Table 1 were mostly variable types. We identified a variable span in each instance and abstracted it to form a variable type, such as [X] *maeni* (before [X]) and [X] *byoukan* (for [X] minutes), where [X] indicates a variable. Each variable can be substituted by a noun (phrase), an adjective, or a numerical value.

The approved and unapproved elements were defined based on the procedures described in section 2.3. As adverbial modifiers can be flexibly inserted into sentences, we also defined a preferred insertion location for each approved element mostly based on the frequency information. The location options are as follows: 1) at the front of the sentence; 2) at the front of the clause; 3) immediately before the verb; 4) after the topical marker は/*wa*; 5) before the nominative case が/*ga*; and 6) after the nominative case が/*ga*.

3.2 Results

The statistics of the constructed lexicon of adverbial modifiers are presented in Table 2 for fixed types and in Table 3 for variable types. In this paper, we use a right arrow '→' to indicate the link from an unapproved element to an approved one.

For the fixed types, we defined 235 approved types and 38 unapproved types, which cover 63,402 and 1,713 tokens in the Corpus-Simple, respectively. This means that about 2.6% (1,713/65,115) of fixed adverbial modifiers in the corpus can be regarded as variations.

For the variable types, we defined 75 approved types and 64 unapproved types, which cover 73,976 and 6,719 tokens in the Corpus-Simple, respectively. As we used the automobile repair manuals as a source for lexicon building, various types of adverbial modifiers regarding *quantity* and *time* were collected, which are indispensable for describing the repair operations.

Category		Example entry	Approved		Unapproved	
Level 1	Level 2		# Type	# Token	# Type	# Token
manner		*yukkurito* →*yukkuri* (slowly), *kinnitsuni*→*kintouni* (uniformly), *kakujitsuni* (securely)	110	21,779	12	660
degree		*ooyoso*→*hobo* (almost), *tashou*→*sukoshi* (slightly), *sarani* (kanji)→*sarani* (kana) (further)	17	2,431	3	92
aspect	proximity	*tadachini*→*suguni* (immediately), *soku*→*suguni* (immediately), *mada* (still)	2	381	3	241
	continuation	*sukoshizutsu*→*jojoni* (gradually), *yuruyakani*→*jojoni* (gradually), *ichijitekini* (temporary)	9	2,937	2	119
	repetition	*futatabi*→*saido* (again), *saido-hajimekara* (again from the beginning)	2	3,971	1	143
	order	*ittan* (kanji)→*ittan* (kana) (for a while), *sonoatoni*→*sonogo* (then), *mazuwa*→*mazu* (first)	21	5,145	5	104
	frequency	*jouji*→*tsuneni* (always), *taezu*→*tsuneni* (always), *teikitekini* (at regular intervals)	6	2,100	2	37
emphasis		*kanarazu* (surely), *kesshite* (never), *zettai*→*zettai-ni* (never)	3	18,475	1	132
others		*suubyou*→*suubyoukan* (for several seconds), *tochude* (halfway through), *sorezore* (respectively)	65	6,183	9	185
		Total	235	63,402	38	1,713

Table 2: Statistics of the lexicon of adverbial modifiers (fixed type)

Category	Example entry	Approved		Unapproved	
		# Type	# Token	# Type	# Token
manner	[X] *to douyou ni* (in the same way as [X]), [X] *to issho ni* (with [X]), [X] *chokuzen de* (just before [X])	3	605	0	0
quantity	[X] *byou ijou* → *sukunakutomo* [X] *byoukan* (for at least [X] seconds), [X] *fun* → [X] *funkan* (for [X] minutes), [X] t o [X] *kai* ([X] to [X] times)	37	16,953	54	6,601
time	[X] *chokugo wa* (immediately after [X]), [X] *ji nado de* → [X] *ji nado ni* (when [X]), [X] *chu de* →[X] *chu ni* (during [X])	31	54,077	10	118
order	[X] *kara jun ni* (from [X] in order), [X] *no* [Y] *jun ni* (in [Y] order of the [X])	4	2,341	0	0
	Total	75	73,976	64	6,719

Table 3: Statistics of the lexicon of adverbial modifiers (variable type)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/
Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society.
Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

4. Lexicon of sentence-ending functions

The sentence-ending functions are word sequences that can be attached to the main verbs for adding various functional information (Kato/Miyata/Sato 2020b). For example, the phrase 表示されない場合がある/*hyoujisa*-**re nai baai ga aru** (**may not be** display**ed**) has three categories of functions—passive voice (**re**), negation (**nai**), and possibility modality (**baai ga aru**). In terms of controlled writing, each function should be expressed in the same form; for example, the possibility modality should be *baai ga aru* instead of *koto ga aru.*

4.1 Building procedure

As mentioned in section 2.2, the unit of a function does not necessarily correspond to a single word; for example, the possibility modality *baai ga aru* is a combination of a noun, particle, and verb. To identify appropriate spans of functions, we used a Japanese sentence-ending analyser Panzer (Sano/Miyata/Sato 2020), which is based on a domain-specific language for Japanese sentence composition (Sato 2020). We first analysed all the sentences in the Corpus-Simple using Panzer and obtained a set of sequences of sentence-ending functions. We then decomposed the 150 most frequent sequences to minimal units of functions and identified 31 function types. These function types were categorised, and approved/unapproved functions were defined based on the method presented in section 2.3.

4.2 Results

Table 4 shows the statistics of the constructed lexicon of sentence-ending functions. The category level 2 can be regarded as the parameters for the category level 1. For example, the voice can be specified by selecting one of the options: active,[2] passive, and causative. Importantly, for most of the categories in level 2, only a single approved element is defined, such as *koto ga dekiru* for potential modality. One of the exceptions is the possibility modality, which includes three approved elements: *baai ga aru*, *kanousei ga aru*, and *osore ga aru.* The examples from the corpus are presented below with their English translations:

(4) 待ち時間が発生する場合がある / *machijikan ga hassei suru baai ga aru*
(Translation: waiting time may be required)

(5) ダイアグを出力する可能性がある / *Daiagu o shutsuryokusuru kanousei ga aru*
(Translation: a DTC may be output)

(6) 不具合が発生するおそれがある / *fuguai ga hasseisuru osore ga aru*
(Translation: this may cause a malfunction)

As the three elements are translated into the same functional word 'may' in English, it is possible to unify them into a single approved form (e.g., *baai ga aru*). Nevertheless, since each expression has a unique nuance, i.e., 'there is a case' (*baai ga aru*), 'there is a possibility' (*kanousei ga aru*), and 'there is a risk' (*osore ga aru*), we retain them as approved elements for the sake of the expressivity of the controlled language.

[2] Although the active voice is an unmarked element in Japanese, we included it in Table 4 to explicitly show the distribution of voice types. The same applies to the affirmative polarity.

Category		Elements	Approved		Unapproved	
Level 1	Level 2		# Type	# Token	# Type	# Token
voice	active	[unmarked]	1	1,384,844	0	0
	passive	*reru*/*rareru*	1	43,537	0	0
	causative	*seru*/*saseru*	1	49,651	0	0
polarity	affirmative	[unmarked]	1	1,360,063	0	0
	negative	*nai*	1	68,318	0	0
modality	obligation	*hitsuyou ga aru*, *nakereba naranai* ⟶ *koto*	2	26,502	1	155
	potential	*koto ga dekiru*	1	6,651	0	0
	possibility	*baai ga aru*, *kanousei ga aru*, *koto ga aru* ⟶ *baai ga aru*, *osore ga aru* (kanji) ⟶ *osore ga aru* (kana)	3	32,130	2	5,490
	tendency	*yasui*	1	831	0	0
	trial	*te-miru*	1	280	0	0
	interrogative	*ka* ⟶ [remove]	0	0	1	198
	determination	*koto ni naru*	1	133	0	0
	permission	*te-yoi*	1	108	0	0
	request	*te-kudasai* ⟶ *koto* (obligation)	0	0	1	74
aspect	state	*te-aru* ⟶ *te-iru*	1	151,202	1	616
	perfect	*te-shimau*	1	1,570	0	0
	preparation	*te-oku*	1	3,828	0	0
	continuation	*te-iku* ⟶ [remove]	0	0	1	214
change		*naru*, *you ni suru*	2	5,093	0	0
completion		*ta*	1	1,963	0	0
parallel	illustration	*tari-suru*	1	1,100	0	0
honorifics	polite	*masu* ⟶ [remove]	0	0	1	495
	humble	*te-itadaku*	1	233	0	0
emphasis		*mo*	1	374	0	0

Table 4: Statistics of the lexicon of sentence-ending functions

5. Lexicon of connectives

In this study, we defined a *connective* as an expression that is located in between clauses (or sentences) to indicate their relationship.[3] We can distinguish two types of connectives: those located at the beginning of the clause (henceforth, clause-beginning connectives); and those located at the end of the clause (henceforth, clause-ending connectives).

[3] Although connectives are also used to combine various elements other than clauses, such as nouns and verbs, in this paper, we focus on clause-level connections.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

5.1 Building procedure

Linguistically, the clause-beginning connectives generally correspond to conjunctions and conjunctive adverbs. We thus used the Japanese morphological analysis tool Juman++ to widely extract these parts of speech as candidates of clause-beginning connectives.[4]

To collect candidates of clause-ending connectives, we identified coordinate and subordinate clauses in the Corpus-Org using Juman++ and KNP and extracted the connectives. As we discovered that the coordinate or adverbial clauses directly identified by the tools are not sufficient for our purpose, we added the pattern of an attributive clause with an adverbial parent element whose attributive relation is 'external' (Teramura 1975–1978).

We manually excluded the irrelevant candidates and controlled variations to define approved and unapproved elements based on the procedures described in section 2.3.

5.2 Results

Table 5 shows the statistics of clause-beginning types of connectives, while Table 6 shows that of clause-ending types. These types are categorised partly based on the typology of Japanese conjunctions by Ishiguro (2016).

It is notable that 37% of clause-ending connective tokens were deemed unapproved variations. The following types of variations were identified and controlled.

- Synonyms: e.g., よって/*yotte* → したがって/*shitagatte* (therefore)
- Character variations: e.g., 時/*toki* (kanji) → とき/*toki* (kana) (when)
- Post-positional particle variations: e.g., 場合/*baai* → 場合は/*baai wa* (in case)

Category	Example	Approved		Unapproved	
		# Type	# Token	# Type	# Token
resultative	*shitagatte* (kana) → *shitagatte* (kanji) (therefore), *yotte* → *shitagatte* (kanji) (therefore), *konotame* (for this reason)	4	625	2	135
adversative	*gyakuni* (conversely), *shikashi* (but)	2	96	0	0
parallel	*mata* (kanji) → *mata* (kana) (also), *katsu* (kana) → *katsu* (kanji) (besides), *sarani* (in addition)	6	4,915	2	23
contrast	*matawa* (kanji)→*matawa* (kana) (or), *aruiwa*→*matawa* (or), *moshikuwa*→*matawa* (or)	1	124	3	40
paraphrase	*tsumari* (in other words)	1	49	0	0
example	*tatoeba* (kana) → *tatoeba* (kanji) (for example)	1	31	1	2
addition	*nao* (in addition), *tadashi* (kanji) → *tadashi* (kana) (however)	2	609	1	398
	Total	17	6,449	9	598

Table 5: Statistics of the lexicon of connectives (clause-beginning types)

[4] This process was conducted in parallel with the lexicon building process for adverbial modifiers described in section 3.1 because the target linguistic elements overlap with each other.

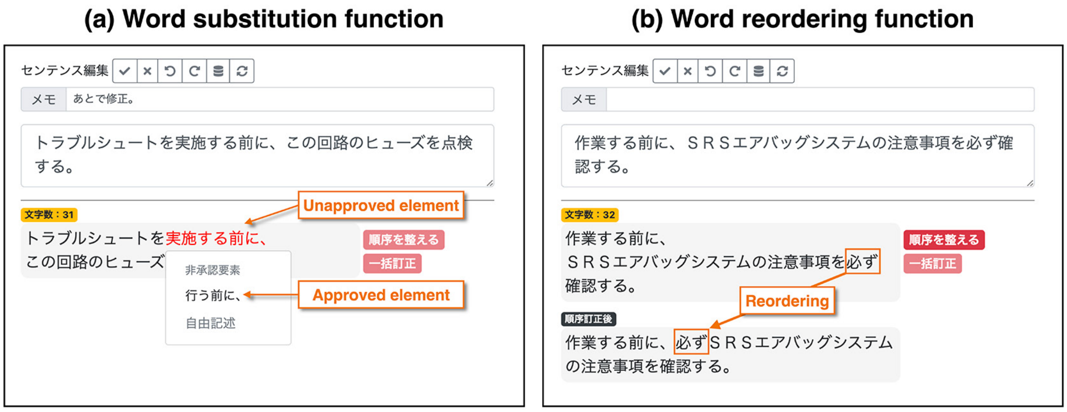
Category			Example	Approved		Unapproved	
Level 1	**Level 2**	**Level 3**		**# Type**	**# Token**	**# Type**	**# Token**
coordinate	resultative	and	V-*te*→V (continuative form)	1	218,165	1	199,945
	illustration	such as	*tari, nado*	2	6,737	0	0
	contradictory	but	*ga*	1	4,912	0	0
	cumulation	in addition	*ue*	1	17	0	0
subordinate	time	when	*toki* (kanji)/*toki wa/sai/sai wa/ jiten de*→*toki* (kana), *sai ni/ toki ni* (kanji)→*toki ni* (kana), *sai niwa, toki niwa*→*toki* (kana)/*toki ni* (kana)	2	15,882	11	21,488
		each time	*tabi ni/goto ni* (kanji)→*goto ni* (kana)	1	115	2	63
		with	*to tomo ni* (kana)→*to tomo ni* (kanji)	1	222	1	119
		before	*mae niwa*→*mae ni, mae wa*	2	17,668	1	80
		after	*nochi/nochi ni/ato* (kana)/*ato ni/ato de* (kanji)→*ato* (kanji), *ato wa* (kanji), *kara*	3	23,563	6	1,849
		then	*ue de* (kana)→*ue de* (kanji)	1	396	1	215
		while	*aida wa, uchi ni*	2	1,623	0	0
		until	*made*	1	9,192	0	0
	condition	in case	*baai ni*→*baai, baaini wa/toki/ toki wa/toki ni/toki niwa/sai wa/sai niwa*→*baai/baai wa*	2	76,611	2	5,397
		if	*ba/naraba/nonara/ nodeareba*→*baai/baai wa, tara*→*ato* (kanji)/*toki* (kana)/ *baai/baai wa*	1	32,203	5	2,449
		if it seems	*you nara/you deareba*→*baai/ baai wa*	0	0	2	403
		though	*mo*	1	4,435	0	0
		as long as	*kagiri* (kana)→*kagiri* (kanji)	1	269	1	9
	methond	by	V (continuative form)→V-*te/ koto de*	1	7,019	0	0
	attendant circumstances	with -ing	*tsutsu*→*nagara/mama*	2	9,597	1	27
		without -ing	*zuni*	1	3,337	0	0
	state	in the state	V (continuative form)→V-*te/ joutai de*	1	7,364	0	0
	purpose	in order to	*you/you ni* (kanji)→*you ni* (kana), *tame* (kanji)/*tame ni*→*tame* (kana), *tame niwa*→*niwa*	2	10,083	6	3,375
	cause	because	*tame*→*node*	1	3,100	2	37,507
		(as a result)	*kekka*→V (end-form)	0	0	1	196

Category			Example	Approved		Unapproved	
Level 1	Level 2	Level 3		# Type	# Token	# Type	# Token
	contradictory	but	*noni*	1	243	0	0
	extent	to the extent	*teido*⟶*teido ni,* *hodo* (kana)⟶*hodo* (kanji)	2	208	2	62
	restrict	just	*dakede*	1	28	0	0
			Total	37	452,989	43	273,427

Table 6: Statistics of the lexicon of connectives (clause-ending types)

6. Writing assistant tool

Implementing the controlled language components mentioned in sections 3–5, we are developing a writing support tool to help writers and editors compose controlled sentences. Figure 2 presents the prototype interface of our tool. Similar to existing controlled language checkers (e. g., Bernth/Gdaniec 2001; Mitamura et al. 2003; Miyata et al. 2016; Nyberg/Mitamura/Huijsen 2003), our tool detects unapproved/non-preferred elements in the input text, suggests candidates of approved/preferred elements, and corrects the target segment when the user selects one of the candidates.



あとで修正。

トラブルシュートを実施する前に、この回路のヒューズを点検する。

トラブルシュートを実施する前に、
この回路のヒューズ

行う前に、

作業する前に、ＳＲＳエアバッグシステムの注意事項を必ず確認する。

作業する前に、
ＳＲＳエアバッグシステムの注意事項を必ず
確認する。

作業する前に、必ずＳＲＳエアバッグシステム
の注意事項を確認する。

Fig. 2: Prototype interface of our writing assistant tool

The word substitution function in Figure 2 (a) is helpful to reduce the word type variations. This function can be implemented by using the pairs of unapproved and approved elements directly obtained from our controlled language components. However, if the input text includes elements that are not registered in the controlled language, we need to specify which approved elements should be suggested in an ad hoc manner. To search for an approved word that is contextually interchangeable to the unregistered element, it would be effective to use the similarity of word vectors obtained from contextual embedding models, such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019).

The word reordering function in Figure 2 (b) can detect the word order variations and suggest the preferred order, based on the location information prescribed in our controlled language. The example in the figure illustrates the suggestion of the appropriate location of an adverbial modifier 必ず/*kanarazu* (surely), the preferred location of which is defined as 'at the front of the clause'. This function is novel as existing controlled language checkers rarely support the correction of word order variations.

7. Conclusion and outlook

In this paper, we reported our attempt to design and build a Japanese controlled language that is intended to support controlled writing of automotive technical documents. The controlled language defines approved and unapproved lexical elements, extensively covering a) case frames, b) case elements, c) adverbial modifiers, d) sentence-ending functions, and e) connectives. The key principles are that it contains linkages between unapproved and approved elements and that it defines preferred word orders for approved elements. Our controlled language components have been constructed through the comprehensive analysis of a large-scale text corpus of automobile repair manuals. While we assume that our controlled language can widely cover the automotive domain, we plan to verify its applicability to other document types in the domain and continuously expand its coverage.

We assume that our methodology to build a controlled language is generally applicable to other domains outside the automotive domain if a sufficiently large corpus is available. While the sizes of components functional words, i.e., d) sentence-ending functions and e) connectives, can be limited, those of the other components, i.e., a) case frames, b) case elements, and c) adverbial modifiers, can become large. To build a manageable controlled language, it would be useful to first focus on a specific text type and examine the growth of coverage according to the lexicon size (Miyata/Sugino 2020). Although our controlled language is based on Japanese, most of the components can also be defined in other languages. Nevertheless, we should carefully reconsider the design of controlled language components when another language is targeted.

We also introduced our prototype of the writing support tool that partly implements the controlled language components. At this stage, the tool purely exploits the constructed controlled language. A wide variety of writing support tools, or augmented writing tools, have been developed to date (Du et al. 2022; Simonsen 2020; Wanner/Verlinde/Alonso Ramos 2013; Yen et al. 2015). To improve the functionality and interface of our tool, it would be effective to utilise technologies found in these existing tools. While our tool is currently intended for post hoc checking scenarios, providing diagnostic functions, more pre-emptive solutions might be useful, such as suggesting subsequent words (e.g., Chen et al. 2019). In future work, we will fully develop the tool and evaluate its usability in practical work scenarios.

References

ASD (2021): ASD Simplified Technical English. Specification ASD-STE100, Issue 8. http://www.asd-ste100.org (last access: 19-03-2022).

Bernth, A./Gdaniec, C. (2001): MTranslatability. In: Machine Translation 16 (3), pp. 175–218.

Chen, M. X./Lee, B. N./Bansal, G./Cao, Y./Zhang, S./Lu, J./Tsay, J./Wang, Y./Dai, A. M./Chen, Z./Sohn, T./Wu, Y. (2019): Gmail smart compose: Real-time assisted writing. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 4–8 August 2019, Anchorage, Alaska, pp. 2287–2295.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag. *https://doi.org/10.14618/phpy-6r66*

Devlin, J./Chang, M.-W./Lee, K./Toutanova, K. (2019): BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2–7 June 2019, Minneapolis, Minnesota, pp. 4171–4186.

Du, W./Kim, Z. M./Raheja, V./Kumar, D./Kang, D. (2022): Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. In: Proceedings of the 1st Workshop on Intelligent and Interactive Writing Assistants, 26 May 2022, Dublin, Ireland, pp. 96–108.

Ishiguro, K. (2016): *Setsuzokushi no gijutsu* [Techniques of Conjunctions].Tokyo. (in Japanese).

Japan Technical Communicators Association (2016): *Nihongo sutairu gaido* [Style Guide for Japanese Documents] (3rd ed.). Tokyo. (in Japanese).

Japio (2018): *Tokkyo raitingu manyuaru: sangyo nihogo* [Patent writing manual: Technical Japanese] (2nd ed.). Tokyo. (in Japanese).

Kato, T./Miyata, R/Tatsumi, M./Sato, S. (2020a): Designing and implementing a support system for simplifying expository text on Japanese cultural assets. In: Proceedings of the 34th Annual Conference of the Japanese Society for Artificial Intelligence, 9–12 June 2020, Online, pp. 1–4. (in Japanese).

Kato, T./Miyata, R./Sato, S. (2020b): BERT-based simplification of Japanese sentence-ending predicates in descriptive text. In: Proceedings of the 13th International Conference on Natural Language Generation, 12–15 December 2020, Online, pp. 242–251.

Kawahara, D./Kurohashi, S. (2006): A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 4–9 June 2006, New York, pp. 176–183.

Kittredge, R. (2003): Sublanguages and controlled languages. In: Mitkov, R. (ed.): Oxford Handbook of Computational Linguistics, Oxford/New York, pp. 430–437.

Kuhn, T. (2014): A survey and classification of controlled natural languages. In: Computational Linguistics 40 (1), pp. 121–170.

Mitamura, T./Baker, K./Nyberg, E./Svoboda, D. (2003): Diagnostics for interactive controlled language checking. In: Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, May 2003, Dublin, Ireland, pp. 237–244.

Miyata, R./Hartley, A./Paris, C./Kageura, K. (2016): Evaluating and implementing a controlled language checker. In: Proceedings of the 6th International Workshop on Controlled Language Applications, 28 May 2016, Portorož, Slovenia, pp. 30–35.

Miyata, R./Sugino, H. (2020): Building a controlled lexicon for authoring automotive technical documents. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion (Volume 1), 7–9 September 2021, Online, pp.171–180.

Nyberg, E./Mitamura, T./Huijsen, W.-O. (2003): Controlled language for authoring and translation. In: Somers, H. (ed.): Computers and the translator. Amsterdam, pp. 245–281.

Peters, M. E./Neumann, M./Iyyer, M./Gardner, M./Clark, C./Lee, K./Zettlemoyer, L. (2018): Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1–6 June 2018, New Orleans, Louisiana, pp. 2227–2237.

Sano, M./Sato, S./Miyata, R. (2020): Detection of functional expressions in Japanese sentence-ending predicative phrases and its application to estimation of rhetorical relation between sentences. In: Proceedings of the 26th Annual Meeting of the Association for Natural Language Processing, 17–20 March 2020, Online, pp. 1483–1486. (in Japanese).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag. *https://doi.org/10.14618/phpy-6r66*

Sato, S. (2020): HaoriBricks3: A domain-specific language for Japanese sentence composition. In: Journal of Natural Language Processing 27 (2), pp. 411–444. (in Japanese).

Simonsen, H. K. (2020): Augmented writing and lexicography: A symbiotic relationship? In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion (Volume 1), 7–9 September 2021, Online, pp. 509–514.

Teramura, H. (1975–78): *Rentaishushoku no shintakusu to imi 1–4* [Syntax and meaning of attributive modification 1–4]. In: Teramura, H. (1992): *Teramura Hideo ronbunshuu 1* [Collection of Papers by Hideo Teramura 1]. Tokyo. (in Japanese).

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, 2–4 November 2018, Brussels, Belgium, pp. 54–59.

Wanner, L./Verlinde, S./Alonso Ramos, M. (2013): Writing assistants and automatic lexical error correction: Word combinatorics. In: Kosem, I./Kallas, J./Gantar, P./Krek, S./Langemets, M./Tuulik, M. (eds.): Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia, pp. 472–487.

Warburton, K. (2014): Developing lexical resources for controlled authoring purposes. In: Isahara, H./Choi, K.-S./Lee, S./Nam, S. (eds.): Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use, 27 May 2014, Reykjavik, pp. 90–103.

Warburton, K. (2021): The Corporate Terminologist. Amsterdam/Philadelphia.

Yen, T.-H./Wu, J.-C./Chang, J./Boisson, J./Chang, J. (2015): WriteAhead: Mining grammar patterns in corpora for assisted writing. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations, 26–31 July 2015, Beijing, China, pp. 139–144.

Contact information

Rei Miyata
Nagoya University
miyata.rei.f2@f.mail.nagoya-u.ac.jp

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 19K20628 and 19H05660, and by the the Japan Science and Technology Agency (JST)-Mirai Program (JPMJMI19G8). The automobile manuals used in this study were provided by Toyota Motor Corporation.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/
Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society.
Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66