6 Adjectives on –i in Japanese language corpora: Distribution, patterns and lexical constraints[1]

Irena SRDANOVIĆ
Juraj Dobrila University of Pula

Abstract
This paper explores Japanese i-adjectives using empirical methods of corpus linguistics and employing state-of-art language resources and a lexical profiling tool. Firstly, this research presents resources that have been used in the analysis and explains their relevance and characteristics. These resources are used to examine the distribution of i-adjectives in the large-scale corpora of contemporary written Japanese, which clarifies which i-adjectives predominate in the overall usage of i-adjectives and how some i-adjectives and adjectival suffixes are more productive than others. Next, this research analyses the distribution of the patterns of the three major roles of adjectives and shows the different tendencies in the usage of their roles and patterns among adjectives. The research focuses on i-adjectives in their attributive role preceding a modified noun and reveals their complexity of patterns and the need to further subcategorize the types of attributive role adjectives have. Furthermore, this study examines lexical constraints in the attributive role of i-adjectives, while discovering some adjectives with no or a rare attributive role.

Keywords: i-adjectives, corpora, distribution, patterns, constraints

1 Introduction

The computer-assisted systematic research of carefully collected large-scale authentic data confirms that lexical items retrieved for utterance constrain the syntactic structure that can be employed for their construction (cf. Schönefeld 1999: 138-9, Stefanowitsch and Gries 2003: 209-10). This empirically based confirmation is one of the main achievements of corpus linguistic research in the analysis of human spoken and written discourse and reminds us of the necessity of employing this methodology to further analyse particular languages and human language in general. Accordingly, the aim of this research is to explore the Japanese language patterns of i-adjectives, with special focus on their role as modifiers of nouns, using the

[1] The previous version of the study "Distribution, semantic and syntactic profile of Japanese i-adjectives" was presented at the conference "XXVII es Journées de Linguistique d'Asie Orientale" in Paris, 2014.

empirical methods of corpus linguistics and employing the latest language resources and lexical profiling tools.

The first part of this paper examines the distribution of i-adjectives in present-day large scale written corpora, differentiating between very frequent and vary rare adjectives and their role in Japanese language productivity. This research touches upon the most productive adjectival suffixes in Japanese which form compound and derived adjectives.

The second part of the paper analyses the distribution of the patterns of the three major roles of adjectives (predicative, attributive and adverbial) as recognized in previous studies (cf. Suzuki 1972, Nishio 1972, Hashimoto and Aoyama 1992) and explores how tendencies of the usage of their roles differ among adjectives. The usage of large-scale corpora, BCCWJ and JpTenTen, and the user-friendly tool Sketch Engine (Kilgarriff et al. 2004, Srdanović et al. 2008, 2013) enable a more thorough exploration of the patterns in their use and some possible constraints. This study uses several adjective examples in order to explore in detail the attributive role of i-adjectives preceding a noun.

This paper is structured as follows: Section 2 explains the resources used in analysis, Section 3 presents the results of the analysis of the distribution of adjectives and their productivity, and Section 4 describes the analysis and results of different patterns and their constraints.

2 BCCWJ, JpTenTen and Sketch Engine: resources used in analysis

This section introduces two large-scale Japanese language corpora, BCCWJ and JpTenTen, and the state of the art corpus query system Sketch Engine. These corpora and the tool are used in this analysis.

Like the British National Corpus (BNC) and many other national corpora, the Balanced Corpus of Contemporary Written Japanese (BCCWJ) was compiled with the aim to be as balanced as possible and with a size of around 100 million words. As described by Maekawa et al. (2013) the originality of this corpus is in its sampling policy. The first two sub-corpora: Publication sub-corpus (PSC) and Library sub-corpus (LSC) used the technique of stratified random sampling. PSC represents the actual state of the publication in Japan and reflects the aspects of the "production" of written texts from books, magazines, and newspapers published between 2001 and 2005. LSC reflects the "circulation" aspect of written Japanese, covering a wide range of the books published between 1986 and 2005 and registered in more than 13 public libraries in the Tokyo metropolis. The third sub-corpus, the Special-purpose sub-corpus (SSC), included sampled data of some specific types of texts, such as public data distributed by the Japanese government, best-selling books, blog data etc.

Thus far there have been various approaches in Japanese word identification, since the Japanese language has no explicit word boundaries and it poses great challenges when identifying words for language processing and corpus analysis. The BCCWJ corpus project decided to employ a two-fold morphological annotation approach with short-unit word (SUW) and long-unit word (LUW) identification (Ogura et al. 2011) using the electronic dictionary UniDic and the morphological annotation tool MeCab. This new approach also covers various orthographic variations under the same units.

The following two examples show how Japanese words are identified using short- and long-unit word annotation. While some words such as particles and single-morpheme units are marked as one unit in both annotation systems (e.g. 私 *watashi* ‘I’, は *wa* ‘[theme particle]’, で *de* ‘[place particle]’), some words are divided into more units in SUW annotation (e.g. 日本/語 *Nihon/go* ’Japanese language [Japanese/language], 勉強/し/て/いる *benkyō/shi/te/iru* ‘study [study_N/suru_AuxV_base/suru_AuxV_renyō/iru_AuxV]]” and combined into one or several units in LUW annotation (e.g. 日本語 *Nihongo* ’Japanese language’, 勉強/し/ている *benkyō/shi/teiru* ‘study [study_N/suru_AuxV_base/teiru_AuxV]). The adjectives on *–i* are typically divided into more units in SUW annotation in case of compound and derived adjectives, such as 興味/深い *kyōmi/bukai* ‘interesting [interest_N/deep_Ai]’, and combined into one unit in LUW annotation 興味深い *kyōmibukai* ‘interesting’. The ordinary simple adjectives are identified as one unit in both SUW and LUW annotation (e.g. 新しい *atarashii* ‘new’).

Example 1-1: Two-fold annotation system (short- and long-unit words)
SUW: /私/は/プーラ/大学/で/日本/語/を/勉強/し/て/いる/。
Watashi/wa/Puura/daigaku/de/Nihon/go/wo/benkyō/shi/te/iru/.
‘I study Japanese at Pula University.’
LUW: /私/は/プーラ大学/で/日本語/を/勉強し/ている/。
Watashi/wa/Puuradaigaku/de/Nihongo/wo/benkyō/shi/teiru/.
‘I study Japanese at Pula University.’

Example 1-2: Two-fold annotation system (short- and long-unit words) applied on i-adjectives

SUW:	新しい	興味/深い
	atarashii	*kyōmi/bukai*
LUW:	新しい	興味深い
	atarashii	*kyōmibukai*
	‘new’	‘interesting’

The next generation of compiled corpora, following the large-scale balanced national corpora such as BNC and BCCWJ, aimed at obtaining corpora much larger in size by using various methodologies in order to collect data from the web and had also achieved a relatively good balance of the data (c.f. Baroni and Bernardini 2004; Baroni and Ueyama 2006; Baroni and Kilgarriff 2006; Sharoff 2006). The first web corpora were close to the national corpora size (c.f. Srdanović et al. (2008) as an example of such Japanese language web data) but a decade later super large-scale corpora were created. For example, the Corpus factory project (Kilgarriff et al. 2010) provided a number of such corpora, including the 10-billion-word Japanese language corpus JpTenTen (Pomikalek and Suchomel 2012; Srdanović et al. 2013), which we will use in the analysis. This corpus uses the same tools for word identification as BCCWJ: MeCab and UniDic with a two-fold annotation system (SUW and LUW).

Finally, we will use the corpus query and lexical profiling system Sketch Engine (Kilgarriff et al. 2004) to search through the two large-scale corpora. The advantages of this tool are in various search possibilities including the advanced functionality Word Sketches that provide a detailed summary of keyword patterns and collocations. In addition, when required, we use the corpus query systems Chunagon to search for data in BCCWJ sub-corpora.

3 Distribution of i-adjectives

3.1 SUW and LUW i-adjectives in BCCWJ

Joyce and Hodošček (2012) report an "approximately thirteen-fold increase in the number of LUW lemma types over the number of SUW lemma types" in BCCWJ. The study on SUW and LUW lemma types for i-adjectives in BCCWJ data (Srdanović 2013a) revealed that there are 761 adjectives annotated as SUW lemma types and 12585 adjectives annotated as LUW lemma types (Table 1). The analysis include general i-adjectives (annotated as Ai.g, jap. 形容詞-一般 *keiyōshi ippan*, such as *yoi*, *tanoshii*, *fukai*), bound adjectives (annotated as Ai.bnd, jap. 形容詞-非自立可能 *keiyōshi-hijiritsukanō*, such as *nai*, *yoi*) and adjectival suffixes (annotated as Suff.ai, jap., 接尾辞-形容詞的 *setsubiji-keiyōshiteki*, such as -rashii, -*ppoi*).[2] The results show that there is an approximately 15-fold increase in the number of LUW over the SUW lemma types for i-adjectives in this large-scale corpus. The large gap reminds us that much linguistic production is actually achieved through combinations of a limited number of elements, which is widely known as the phenomenon of language economy.

[2] While bound adjectives do not appear as a category in LUW, there are still some adjectival suffixes not attached to any LUW item in the BCCWJ data.

Table 1. I-adjectives annotated as SUW and LUW lemma units in the corpus BCCWJ (based on UniDic + MeCab annotation)

Annotation schema	**Examples of annotated i-adjectives (general and bound adjectives)**	**No. of i-adjective lemma units in JpTenTen**
SUW	無い *nai* 'nonexistent, [indicates negation]', 良い *yoi* 'good', 美味しい *oishii* 'delicious', 楽しい *tanoshii* 'funny', らしい *rashii* 'such as', 易い *yasui* 'simple'	761
LUW	無い *nai* 'nonexistent, [indicates negation]', 良い *yoi* 'good', 美味しい *oishii* 'delicious', 楽しい *tanoshii* 'enjoyable', 人間らしい *ningenrashii* 'human', 色っぽい *iroppoi* 'amorous', 間違い無い *machigainai* 'certain', 分かり易い *wakariyasui* 'easy to understand'	12585

3.2 I-adjectives with high frequency and high coverage

As previously discussed in Srdanović (2013a), the study of i-adjectives annotated as LUW lemma units in BCCWJ reveals that highly frequent 25 (24+1) i-adjectives take up 62% (20%+42%) of the overall occurrences of i-adjectives, when calculating their tokens (Table 2). The first 127 i-adjectives on the frequency list take up almost 90% (62%+27%) of the overall occurrences of this part of speech category. We can see that the number of highly frequent adjectives is relatively small and that these adjectives constitute the highest coverage of adjective usage in BCCWJ. On the other hand, there is great number of very rare adjectives that appear only once or rarely. The rare adjectives that appear from 1 to 10 times cover around 87% of overall adjective usage, when calculating types. Figure 1 shows the usage of adjectives in relation to their frequency span in the corpus and the number of i-adjective lemmas that appear in the corpus in the specific frequency span.

It must be noted that highly frequent items that constitute the large amount of coverage in a language are more likely to be encountered by foreign language learners and therefore should be given priority in vocabulary learning. As Nation (2001) already proposed for English vocabulary learning, a basic 2000-3000 words cover around 70-80% of language usage and therefore need to be placed on the priority list for learners. This approach can be applied to other languages and is in line with the results of the distribution of i-adjectives, where the first group of adjectives takes up 62% of overall usage. The next group of 102 i-adjectives is also very widely used (27%) and, therefore, could be covered gradually after the first group of i-adjectives. All in all, both groups of i-adjectives cover 89% of i-adjective usage.

Table 2. The number of LUW i-adjective lemma tokens and types per frequency span in BCCWJ

Freq span of Ai	LUW Adj (token)	%
100000～	279682	20,02
10000～100000	591458	42,33
1000～10000	378934	27,12
100～1000	94210	6,74
10～100	33739	2,41
1～10	19122	1,37
Total	1397145	100

Freq span of Ai	LUW Adj (type)	%
100000～	1	0,01
10000～100000	24	0,19
1000～10000	102	0,81
100～1000	341	2,71
10～100	1064	8,45
1～10	11053	87,83
Total	12585	100

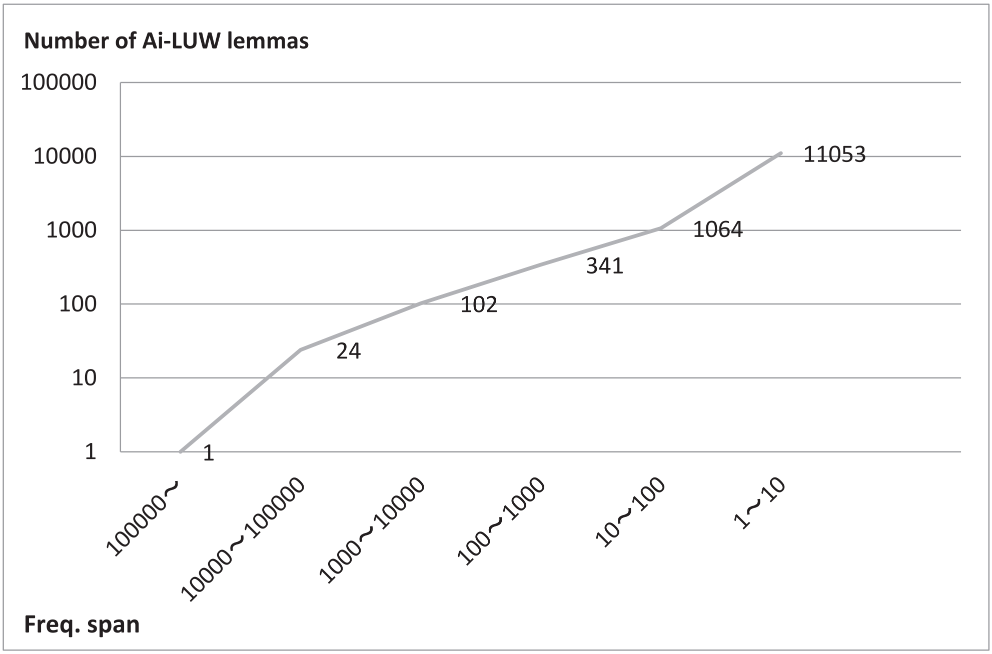


Figure 1. I-adjective LUW lemma appearance in BCCWJ based on frequency span (type-token ratio)

What follows is a list of the 25 most frequent i-adjectives (annotated as LUW lemma) that appear in BCCWJ.[3]

無い *nai* 'non-existent, [indicates negation]', 良い *yoi* 'good, [indicates permission]', 良い *yoi* 'good', 多い *ōi* 'many', 高い *takai* 'high', 大きい *ōkii* 'big', 悪い *warui* 'bad', 強い *tsuyoi* 'strong', 新しい *atarashii* 'new', 早い *hayai* 'fast', 長い *nagai* 'long', 少ない *sukunai* 'few', 旨い *umai* 'skillful', 若い *wakai* 'young', 凄い *sugoi* 'amazing', 小さい *chīsai* 'small', 難しい *muzukashii* 'difficult', 楽しい *tanoshii* 'joyful', 深い *fukai* 'deep', 美味しい *oishii* 'tasteful', 近い *chikai* 'close', 低い *hikui* 'low', 嬉しい *ureshii* 'happy', 面白い *omoshiroi* 'interesting', 広い *hiroi* 'wide'

3.3 Low frequency adjectives, suffixes and their productivity

Interestingly, there is a very high number of low frequency adjectives – more than 11000 adjectives that appear only once or up to ten times. Carefully observing their production in the corpus examples, we notice a large number of compound and derived adjectives that are formed using various suffixes, which indicates the capacity of i-adjectives for productivity and creativity in Japanese. Table 3 shows the list of all suffixes that are used to produce more than a hundred various i-adjectives in the corpus. The second line shows the number of different adjectives created using a specific suffix, where the frequency of each adjective is only one. The third line shows the total number of different adjectives that are created using a specific suffix. Finally, the fourth line shows the total number of all the appearances of adjectives with a specific suffix. The most productive suffixes are *-rashii*, *-yasui*, *-ppoi*, *-gatai/nikui.*[4] However, different kinds of productivity can be noticed: *-ppoi/-poi* seems to have the most tendencies for variety and creativity as the number of adjective that appear only once is quite high, in comparison to the number of different adjectives and all their appearances. *-rashii* and *–tsurai* have similar tendencies, so we might expect more derived adjectives in the case of these suffixes. On the other hand, *-nai, -yoi/ii, -fukai/bukai* have a limited number of different adjectives that are repeated quite often.

[3] This list also includes the most frequent function words *nai* (無い) and *yoi/ii* (良い). Note that the word *tai* (たい) is annotated as an auxiliary verb. Also, note that there are some slight differences among the list of 25 of the most frequent i-adjectives in SUW and LUW annotation data – 欲しい *hoshii* 'to want [general i-adjective]' and 易い *yasui* 'easy; likely to …[general i-adjective and adjectival suffix]' appear higher in the frequency list in data annotated as SUW.

[4] Refer to Srdanović (2013) for more details on the compounding of adjectives with their suffixes, most frequent compounds, i-adjectives with the suffix *-kusai*, and for a discussion on the implications of the results on Japanese language learners.

Table 3. Suffixes and their productivity in producing i-adjectives

Suffix	**No of Ai (freq=1)**	**No of Ai (diff)**	**No of Ai (total)**
らしい *rashii* 'seemingly/like'	2376	2995	15734
易い *yasui* 'easy'	1213	2316	22738
っぽい *ppoi* 'like'	1031	1326	5694
難い *nikui/gatai* 'hard'	939	1718	15158
無い *nai* 'non-existent [negation]'	458	632	18540
辛い *tsurai* 'painful/tough'	203	343	1650
良い *yoi* 'good'	194	245	5200
臭い *kusai* 'smelly/-ish'	181	292	3796
ぽい *poi* 'like'	158	185	238
深い *fukai/bukai* 'deep'	124	210	4705

When closely observing the results from the viewpoint of Japanese language education, it can be noticed that it would be very effective to provide learners with detailed information on the process of suffix productivity in order to provide support during the process of learning how to produce new, correct and meaningful adjective combinations, especially in cases of highly productive suffixes. On the other hand, restricted and lexicalized adjectives can be introduced and learned one by one or in meaningful groups in order to make learning maximally efficient.

4 I-adjective patterns and their constraints

This section explores the distribution of i-adjective patterns in Japanese language corpora and their constraints in the attributive role.

4.1 Distribution of patterns for i-adjectives: case of *takai*

This section takes the Japanese i-adjective *takai* 'high, tall, expensive' as an example and explores its most frequent patterns that appear in the corpus JpTenTen, using the Sketch Engine Word sketch tool. Figure 2 shows that among the recognized patterns the following appear most frequently:

1) *Takai* as a noun modifier, in its attributive role (*rentai-kei*) preceding a noun (高い+N, *takai* + N 'high + N', 18%), such as 高い山 *takai yama* 'a high mountain';
2) *Takai* as an adjectival predicate (*shūshi-kei*) (Nが高い[concl], *N ga takai* 'N is high', 18%), such as 効果が高い *kōka ga takai* 'the effect is high'; and

3) *Takai* in combination with some suffixes, such as 高さ *takasa* 'height', 高過ぎ *takasugi* 'too high' (18%).

In the annotated corpus data, another type of attributive role (*rentai-kei*) of the adjective *takai* appears as an adjectival predicate in adnominal clauses and is quite frequent:

4) Adjectival predicate in adnominal clause (Nの/が高いN, *N no/ga takai N* 'N with high N', 16%).

This attributive role of *takai* is both a noun modifier preceding a noun but also an adjectival predicate of a preceding relative clause, forming the construction *N ga/no Ai N*. For example, 完成度の高い作品 *kanseido no takai sakuhin* 'work with a high degree of perfection/completion', or 背が高い方 *se ga takai kata 'a tall person [lit. a person with a high back*]'. In this construction, the so called '*ga/no* conversion' (cf. Harada 1971) occurs with the case particles *ga* and *no*. Since this kind of usage is quite prominent in the case of the adjective *takai*, yet also appearing in the usage of some other adjectives (Srdanović 2013b), the author suggests that there is a need to differentiate it from a pure attributive role in a noun phrase, at least in corpus annotation data. This would be possible either by using a narrower tagset for grammatical roles (forming a new inflection tag 'attributive-predicative role') or by improving parsing (syntactic analysis) for these kinds of grammatical constructions. Also, this pattern with *ga/no* conversion needs to be differentiated from the pattern with the possessive *no* particle, where two attributives (nominal and adjectival) appear modifying the same noun (*N1noAiN2*, see the case of adjective *aoi* in Section 4.2).

Besides these usages, there are also:

5) Adnominal clause with adjectival predicate and omitted case particle or a compound of a noun and *takai* (10%), such as テンション高い*tenshon-takai* 'high tension'.
6) Adverbial form preceding a verb (*renyō-kei*) (高く+V, *takaku* + *V* 'V + high_adv)' (9%), such as 高くなる*takaku naru* 'to become expensive'、高く売る*takaku uru* 'to sell at a high price'. For adverbial forms, although considered one of main three forms in adjective usage, the results showed that, in the case of *takai*, it is not as prominent when compared to other forms.
7) Conjunctive form of *takai* used to link adjectives or clauses (*renyō-kei*) (Nが高く(て)[cont], *N ga takaku(te)[cont]* 'N is high and …' (8%).

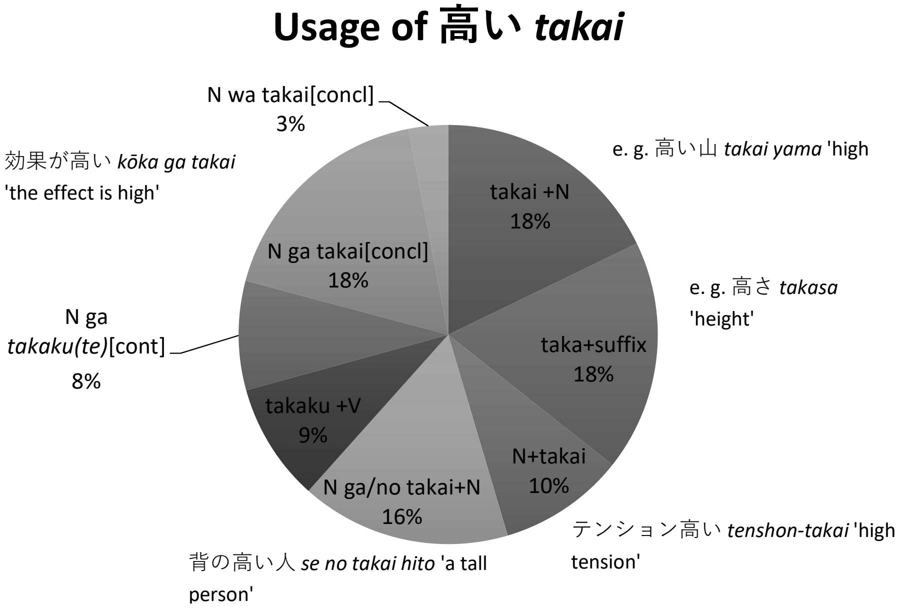


Figure 2. Usage of *takai* in the JpTenTen corpus

4.2 Adjectives in attributive form preceding a modified noun: a previous study revisited

In a previous study (Srdanović 2013b), the author explored the syntactic structure of the attributive role of an adjective preceding a noun (Ai_rentai + N) for a number of adjectives of high, medium and low frequency with an aim to gain better withdrawal results for Japanese adjectives in their attributive form from the SkE collocational functionality Word Sketches. A hundred random examples of *Ai_rentai +N* have been taken from the corpus for each adjective and analysed.

The research findings showed that different adjectives have different tendencies in forming the patterns. For example, the adjectives 寒い *samui* 'cold', 親しい *shitashii* 'close', 甘い *amai* 'sweet' have the tendency to appear mostly in simple noun phrases of an adjective modifying a noun (Ai + N) (90%, 93%, 86% respectively), such as 寒い季節 *samui kisetsu* 'cold+season = cold season'. The adjectives 多い *ooi* 'a lot of' and 高い *takai* 'high' show a very high tendency when forming adnominal clauses with an adjectival predicate. More specifically, there is a large number of *ooi + N* or *takai + N* cases (91% and 54% respectively) where the adjective has the role of a modifier and at the same time the role of an attributive predicate (for example, 雨の多い国 *ame no ōi kuni* 'rain+-CONV_no+lots of+country = a country with lots of rain, or 質の高いサービス *shitsu no takai sābisu* 'quality+CONV_no+high+service = a high-quality service'). Such cases clearly differentiate from other observed patterns, such as simple *Ai* + *N* pattern, where an adjective has the pure role of a noun modifier, and therefore they deserve their own sub-classification. This is especially relevant in the domain of corpus annotation, but might also deserve to be reconsidered within the domain of Japanese language grammar.

Furthermore, there are patterns with two attributives (nominal and adjectival) modifying the same noun (*N1noAiN2*), that are more characteristic for adjectives such as青い *aoi* 'blue' or 甘辛い *amakarai* 'salty-sweet'. In this pattern, the particle *no* connects *N1* to *N2* and the i-adjective modifies *N2*, so that the order of modification is *N1no (Ai+N2)* (17% and 12% respectively). For example, 日本の青い空 *Nihon no aoi sora* 'Japan + POSS_no +blue+sky = Japanese blue sky'. This kind of pattern needs to be differentiated from adnominal clauses with an adjectival predicate, where the particle *no* appears as, the so called '*ga/no* conversion' (cf. Harada 1971), which we will explore further in Section 4.3.

There are also a few cases of adnominal clauses with adjectival predicate and omitted case particle or compounds of a noun and an adjective, also in the case of adjectives such as 多い *ōi* 'a lot of' and 高い *takai* 'high' (2% and 5%), e. g. 香り高いコーヒー *kaoritakai kōhī* 'aroma+high+coffee = aromatic coffee'.

The above results of the corpus-based study, including the results from Figure 2, show that there is a need to reconsider corpora annotation but also the traditional way of observing the syntax of adjectives and there usage in the Japanese language. The three major roles of adjectives: predicative, attributive and adverbial (cf. Suzuki 1972, Nishio 1972, Hashimoto and Aoyama 1992) need to be reconsidered for their subgroups and further explored based on the large-scale corpora available for contemporary Japanese to widen the understanding of their behaviour and the patterns of adjectives.

4.3 Exploring the pattern 'N no/ga takai N'

This section takes the patterns Nが高いN *N ga takai N* 'N with high N' and Nの高いN" (*N no takai N* 'N with high N') as an example and shows the results of the quantitative corpus-based analysis of these patterns. Table 4a and 4b display the reoccurrences of various combinations within the pattern. For example, "~度が高いこと/もの/作品" (… *do ga takai koto/mono/sakuhin* 'thing/work with high level of ...') in Table 4a, 質の高いサービス (*shitsu no takai sābisu* 'service of a high quality') in Table 4b, or 背の/が高い人・方 (*se no/ga takai hito/kata* 'a tall person [lit. a person with a high back]' both Tables 4a and 4b).

The *ga/no* conversion phenomenon, where the particle *ga* stands for 'Nominative Case' and *no* for 'Genitive Case', has been explored and explained in various studies (cf. Harada 1971, Tsujimura 1996: 264-266). These studies describe the possibilities of the conversion in relative clauses in Japanese and state that there is no apparent

Table 4a: Corpus-based analysis of the pattern Nが高いN *N ga takai N* ‘N with high N’)

~が高い~ ga takai	こと koto 'thing, [nominalizer]'	ため tame 'for, because'	もの mono 'thing'	人 hito 'person'	方 kata 'person'	気 ki 'mood, spirit'	場合 baai 'in case of'	ところ tokoro 'place, when'	そう sō 'it seems, it is said'	わけ wake 'reason, it is not so...'	時 toki 'times, when'	せい sei 'back'	事 koto 'thing'	分 bun 'part, point'	作品 sakuhin 'work'	理由 riyū 'reason'	傾向 keikō 'tende ncy'
度 do 'degree'				227	197	155	167	119	53	149	48	61	114	193	205		53
率 ritsu 'rate'			215	107	97	374	144	112	198	112	45	64	158	49		141	199
～度 do 'degree'	971	614	664														
～率 ritsu 'rate '	1409	578															
背 se 'back'	251	82	49	565	311					79		53		56			
レベル reberu 'level'	343	110	144	240	108	63	78	93		74			49				
人気 ninki 'popularity'	250	178	178						179						56	127	
効果 kōka 'effect'	567	172	269						110				52				
値段 nedan 'price'	316	93	219		66	81		52					55	88			
能力 nōryoku 'ability'	306	153	60	260	113					62							
リスク risuku 'risk'	494	180	104	102			72										
評価 hyōka 'evaluation'	189		152	69											124	93	
敷居 shikii 'treshhold'	61		175			145											
確率 kakuritsu 'probability'	256	98	75			119			99	71			50				
標高 hyōkō 'elevation'	61	258						170				102					
気温 kion 'temperature'	56	104									97	121					
方 hou, kata 'more, person'	159	50	60			139	96			56							46
湿度 shitsudo 'humidity'	63	117					58				113	86					
頻度 hindo 'frequency'	205	105	174	65	55		76										
価格 kakaku 'price'	225	107	100														

Table 4b: Corpus-based analysis of the pattern Nの高いN *N no takai* N 'N with high N')

~の高い~ no takai	もの mono 'thing'	人 hito 'person'	作品 sakuhin 'work'	方 kata 'person'	サービス sābisu 'service'	商品 shōhin 'product'	ところ tokoro 'place'	製品 seihin 'product'	情報 jōhō 'information'	選手 senshu 'player'	男 otoko 'man'	女性 josei 'woman'	医療 iryō 'healthcare'	場所 basho 'place'	靴 kutsu 'shoes'
度 do 'degree'		303		220	180	287	131	246	134	87		63	57	118	
質 shitsu 'quality'	727		447		1474	200		170	410	70			933		
背 se 'back'	283	1411		628						141	992	392			
レベル reberu 'level'	587	574	320	163	66		246			250		82			
人気 ninki 'popularity'	266		250		48	519									
～度 do 'degree'	1984		1303												
価値 kachi 'value'	858		73		365	357		323	167						
効果 kōka ''effect	489				47	154		85							
クオリティ kuoriti 'quality'	389		368		92	135		89							
率 ritsu 'rate'	268	53		57		180	102	66						54	
精度 seido 'accuracy'	254							62	185						
能力 nōryoku 'ability'	107	416		243						399					
標高 hyōkō 'elevation'							526							356	
意識 ishiki 'awareness'		521		302						52		161			
湿度 shitsudo 'humidity'							117							220	
リスク risuku 'risk'	150	182		64		57									
頻度 hindo 'frequency'	433														
評価 hyōka 'evaluation'	181	86	263			49				68					
身分 mibun 'status'	70	468		50								85			
ヒール hīru 'heel'															614
品質 hinshitsu 'quality'	182				64	111		118							
カロリー karorī 'calories'	359														

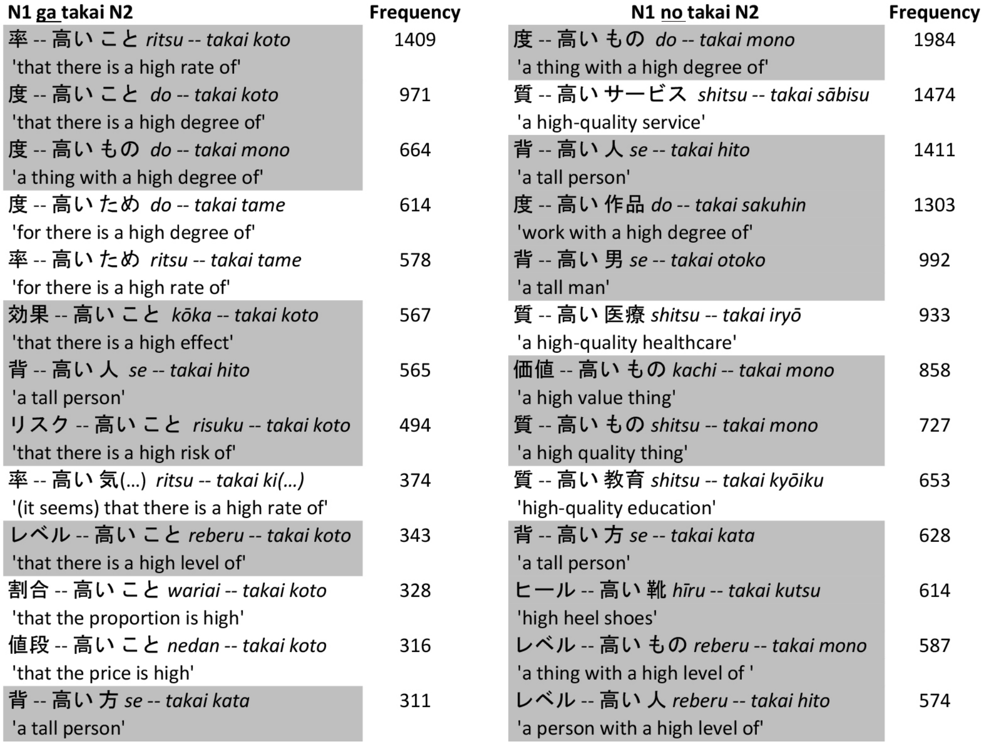
difference in meaning. Looking into the differences in occurrences of the particles *ga* and *no* in the specified patterns throughout the corpus, revealed that there are differences in the distribution between the use of *ga* and *no* in this kind of pattern and that some combinations of words in the pattern tend to appear more or only with one of the particles. The results also revealed that *ga* tends to be more in use in combination with functional words and in an abstract sense (e.g. こと *koto* 'thing, [nominalizer]', ため *tame* 'for, because'). On the other hand, *no* is used twice to three times more in this kind of pattern than *ga*, and tends to appear more in concrete cases (質の高いサービス *shitsu no takai sābisu* 'high level service', ～度の高い作品 *…do no takai sakuhin* 'work with a high degree of…'[5], ヒールの高い靴 *hīru no takai kutsu* 'shoes with high heels' etc.).

The details of the most frequent patterns are also presented in Table 5, where unmarked patterns appear frequently only with one of the case particles, *no* or *ga*, while marked patterns appear in both cases.[6] For example, ～度 -- 高いもの *…do -- takai mono* 'a thing with a high degree of ...' (*ga*: 664 / *no*: 1984) and 背 -- 高い人 *se -- takai hito* 'a tall person' (*ga*: 565 / *no*: 1411) appear with both case particles quite frequently, although we can notice almost three times more usage with *no* in these particular examples as well. There are plenty of examples where combinations have high tendencies and appear more often or only with one case particle. The particle *no* particularly tends to appear more than the particle *ga* within patterns where N1 is 質 *shitsu* 'quality' and N2 are words such as サービス *sābisu* 'service', 医療 *iryō* 'healthcare', 教育 *kyōiku* 'education', 情報 *jōhō* 'information'. On the other hand, the particle *ga* rather appears in the patterns with funcational words ため *tame* 'for, because', こと *koto* 'thing, matter [nominalizer]', such as ～度が高いため *… do ga takai tame* 'for there is a high degree of …', 率が高いため *ritsu ga takai tame* 'for there is a high rate', 割合が高いこと *wariai ga takai koto* 'that the proportion is high'. The reason for this is related to the fact that the conversion from *ga* to *no* following N1 enables another *ga* to appear as a particle following N2 and marking it as a subject of a sentence. Further elaboration on syntactic, semantic or pragmatic reasons for this is required.

[5] 度 *do* 'degree' appears as a suffix which forms words with various nouns. Within the target pattern these nouns are, for example, 完成度 *kanseido* 'degree of perfection; level of completion', or 難易度 *nan'ido* 'degree of difficulty'.

[6] For each pattern, *N1ga takai N2* and *N1 no takai N2,* the most frequent thousand occurances are extracted, thus the patterns of frequency less than 16 for the pattern *N1ga takai N2* and less than 31 for the pattern *N1 no takai N2* are not taken into account.

Table 5. Tendencies in the pattern: particles *ga* and *no*



N1 ga takai N2	Frequency	N1 no takai N2	Frequency
率 -- 高い こと *ritsu -- takai koto* 'that there is a high rate of'	1409	度 -- 高い もの *do -- takai mono* 'a thing with a high degree of'	1984
度 -- 高い こと *do -- takai koto* 'that there is a high degree of'	971	質 -- 高い サービス *shitsu -- takai sābisu* 'a high-quality service'	1474
度 -- 高い もの *do -- takai mono* 'a thing with a high degree of'	664	背 -- 高い 人 *se -- takai hito* 'a tall person'	1411
度 -- 高い ため *do -- takai tame* 'for there is a high degree of'	614	度 -- 高い 作品 *do -- takai sakuhin* 'work with a high degree of'	1303
率 -- 高い ため *ritsu -- takai tame* 'for there is a high rate of'	578	背 -- 高い 男 *se -- takai otoko* 'a tall man'	992
効果 -- 高い こと *kōka -- takai koto* 'that there is a high effect'	567	質 -- 高い 医療 *shitsu -- takai iryō* 'a high-quality healthcare'	933
背 -- 高い 人 *se -- takai hito* 'a tall person'	565	価値 -- 高い もの *kachi -- takai mono* 'a high value thing'	858
リスク -- 高い こと *risuku -- takai koto* 'that there is a high risk of'	494	質 -- 高い もの *shitsu -- takai mono* 'a high quality thing'	727
率 -- 高い 気(…) *ritsu -- takai ki(…)* '(it seems) that there is a high rate of'	374	質 -- 高い 教育 *shitsu -- takai kyōiku* 'high-quality education'	653
レベル -- 高い こと *reberu -- takai koto* 'that there is a high level of'	343	背 -- 高い 方 *se -- takai kata* 'a tall person'	628
割合 -- 高い こと *wariai -- takai koto* 'that the proportion is high'	328	ヒール -- 高い 靴 *hīru -- takai kutsu* 'high heel shoes'	614
値段 -- 高い こと *nedan -- takai koto* 'that the price is high'	316	レベル -- 高い もの *reberu -- takai mono* 'a thing with a high level of '	587
背 -- 高い 方 *se -- takai kata* 'a tall person'	311	レベル -- 高い 人 *reberu -- takai hito* 'a person with a high level of'	574

4.4 Lexical constraints of attributive roles

In order to explore the collocational relations of i-adjectives and nouns, the 500 most frequent adjectives and their most frequent collocates have been selected from JpTenTen and BCCWJ (Srdanović 2014). For highly frequent adjectives up to 100 collocates are taken, for the remainder, up to 50 collocates. To avoid unclear data, collocates below frequency 5 for JpTenTen and 2 for BCCWJ have been excluded (since BCCWJ is smaller in size, set frequency is also lower). As can be noted in Table 6, while 500 of the most frequent adjectives could be retrieved from both corpora without too many problems, the number of discovered collocations significantly lowers in the case of a smaller corpus, which confirms what has been already stated about corpora usage limitations in relation to its size and different language phenomena. This must be considered in descriptive linguistic studies and lexicographic work in order to uncover and describe complete collocational information.

Furthermore, the analysis of the retrieved data on i-adjectives and noun relations revealed that a number of target adjectives discovered in both corpora have no attributive role or a quite rare one when compared to the other usages of a particular adjective. Table 6 reveals a number of adjectives with no or a very rare attributive role. Here we can expectedly notice the opposite trend, i.e., that the smaller corpus lists a larger number of adjectives with no or a rare attributive role (83), while the larger corpus recognized more cases of attributive role usage in some of the adjectives and, thus, lists a much smaller number of adjectives with no or a rare attributive role (23). For example, 止む無い *yamunai* 'unavoidable', listed in BCCWJ as one with no or a rare attributive role (only 3 cases), appears in JpTenTen with an attributive role in more than 200 cases (for example, 止む無い事情 *yamunai jijō* 'unavoidable circumstances', 止む無いこと *yamunai koto* 'unavoidable thing' etc.).

A closer look into the results of some of the i-adjectives implies the need to use larger language data when attempting to discover particular lexical constraints. Thus, larger corpora is more reliable for this purpose.

Table 6. Lexical constraints of attributive roles in i-adjectives observed in two corpora

Corpus	**Number of adjectives**	**Number of collocations**	**No or very rare attributive role**
JpTenTen	500	23220	23
BCCWJ	500	9218	83

Table 7 shows a more detailed corpus analysis in some of the retrieved adjectives with no or a very rare attributive role. For some adjectives, no attributive role is discovered in the corpus and they do not directly proceed and modify a noun, such as the adjectives *tegarui* 'easy' and *tokorosemai* 'crowded'. These adjectives appear, rather, in other forms and patterns: *tegarui* in its nominalized form *tegarusa* 'easiness' or in compound 手軽すぎる *tegarusugiru* 'too easy', and *tokorosemai* in its predicative form preceding a clause, such as *tokorosemashi to narande iru* 'to be lined up crowdedly'.

However, some adjectives do appear in attributive roles but these were not retrieved by the search engine for various reasons (e.g. different orthographic form), for example 訝しい *ibukashii* 'suspicious' with rare attributive role いぶかしい顔 *ibukashii kao*'suspicious face'.

A future task of this study would be to explore some other search methods in order to retrieve more definite data on lexical constraints from the available corpora and to differentiate between non-attributive roles and very rare attributive role.

Table 7. I-adjective with no or very rare attributive role: revisited

Adjectives (Ai)	Freq	Most frequent patterns	Modified noun
tegarui 手軽い 'easy'	14542	手軽さ *tegarusa* 'easiness', お手軽さ *otegarusa* 'easiness', （お）手軽すぎる *(o)tegarusugiru* 'too easy'	/
ibukashii 訝しい 'suspicious'	12271	いぶかしげ な顔 を する *ibukashigena kao wo suru* 'to have a suspicious face', いぶかしく 思う *ibukashiku omou* 'to think suspiciously', 訝しがる *ibukashigaru* 'to be suspicious', いぶかしげ に *ibukashige ni* 'suspiciously'	* with different orthographic form いぶかしい顔 *ibukashii kao* 'suspicious face' (13) ・表情 *hyōjō* 'expression' (11)…
tokorosemai 所狭い 'crowded'	11569	所狭しと 並ん で いる *tokorosemashi to narande iru* 'to be lined up crowdedly', 所狭し と 並べ られ て	/
nikui/gatai 難い 'hard'	9273	し難い *shinikui* 'hard to do', わかり難い *wakarinikui* 'difficult to understand', サビ にくい *sabinikui* 'resistant to rust', 代え 難い すばらしい もの *kaegatai subarashii mono* 'amazing things hard to replace' (？)	* as a compound adjective
omowashii 思わしい 'suitable, well, convenient'	7588	体調 が 思わしく ない *taichō ga omowashiku nai* 'not to feel well / to be in poor health', 結果 が 思わしく なかっ た *kekka ga omowashiku nakatta* 'the results were not favorable' *often used in negation	思わしい 結果 *omowashii kekka* 'faborable results' (41)・効果 *kōka* 'effects' (13) が 出ない *ga denai* 'cannot get'…

5 Conclusion

This research showed the importance of using large-scale language resources and state-of-the-art tools in empirical studies in order to challenge the currently available traditional approaches to language study and existing findings. It explores Japanese i-adjectives from various perspectives, focusing on the distribution of i-adjectives in present-day corpora, their patterns and constraints. The research on the distribution of i-adjectives revealed the i-adjectives that predominate over the usage of other i-adjectives and provided some new insights into the productivity of adjectival suffixes in Japanese. Next, this research explored the distribution of patterns and their adjectival roles in the case of the adjective *takai* 'high' and revisited the previous research on the usage patterns of a number of i-adjectives in their attributive role. The research results indicated a need to reconsider the three major roles of adjectives: predicative, attributive and adverbial for their subcategorization in the domain of corpus annotation but also within the domain of the grammar of the Japanese language.

Furthermore, the patterns Nが 高いN *N ga takai N* 'N with high N' and Nの高いN" (*N no takai N* 'N with high N') are examined and other than providing the most frequent patterns, the study had some interesting finding on the *ga/no* conversion in the specified pattern: e. g. *ga* tends to be used for more abstract phenomena while the more frequent *no* for more concrete things. Finally, this study examined the lexical constraints in the attributive forms of i-adjectives and discovered some adjectives with no or a rare attributive role.

Literature

Baroni, M. and Ueyama, M. (2006) Building general- and special-purpose corpora by Web crawling. *Proceedings of the 13th NIJL International Symposium*, 31-40.

Baroni, M. and Bernardini, S. (2004) Bootcat: Bootstrapping corpora and terms from the web. *Proceedings of LREC*, 1313-1316.

Baroni, M. and Kilgarriff, A. (2006) Large linguistically-processed web corpora for multiple languages. *Proceedings of EACL*, 87-90.

Joyce, T., Hodoščček, B. and Nishina, K. (2012) Orthographic representation and variation within the Japanese writing system: Some corpus-based observations, *Written Language and Literacy*, 15 (2) (Special issue: Units of language - units of writing, edited by T. Joyce and D. Roberts), John Benjamins Publishing Company, 254-278.

Kilgarriff, A., Reddy, S., Pomikálek, J. and PVS, A. (2010) A corpus factory for many languages. *Proceedings of LREC*, 904-910.

Kilgarriff, A., Rychly P., Smrž, P., Tugwell, D. (2004) The Sketch Engine. *Proceedings Euralex*, 105-116.

Harada, S. (1971) Ga-no conversion and idiolectal variations in Japanese. *Gengo Kenkyū (Language Research)* 60, 25-38.

Hashimoto, M. and Aoyama, F. (1992) Three usages of adjectives. *Kēryo Kokugogaku (Mathematical Linguistics)*, 18 (5), 201-214.

Nation, P. (2001) *Learning vocabulary in another language*. Cambridge University Press

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y. (2013) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*. Springer Netherlands

Nishio, T. (1972) *A descriptive study of the meaning and uses of Japanese adjectives. NLRI 44*, Shuei Shuppan

Ogura, H., Koiso, H., Fujiike, Y., Miyauchi, S., Konishi, H., Hara, Y. (2011) *Gendai nihongo kakikotoba kinkō kōpasu keitai-ron jōhō kitei-shū dai-yon-ban*. Technical Report LR-CCG-20-05-01, The National Institute of Japanese Language and Linguistics.

Pomikálek, J. and Suchomel, V. (2012) Efficient web crawling for large text corpora, *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*.

Schönefeld, D. (1999) Corpus Linguistics and Cognitivism. *International Journal of Corpus Linguistics*, 4 (1), 137-171.

Sharoff, S. (2006) Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*. Gedit.

Srdanović, I., Erjavec, T. and Kilgarriff, A. (2008) A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing)* 15 (2), 137-159.

Srdanović, I. (2013a). Japanese i-adjectives as short and long unit words: implications for language learning. *PACLING 2013: Conference proceedings*, September 24, 2013 Tokyo: Pacific Association for Computational Linguistics, 8pp.

Srdanović, I. (2013b) Collocation and Syntax: Adjective and Noun Collocations, *Proceeding of the 4th Japanese corpus linguistics workshop*, Department of Corpus Studies/Center for Corpus Development, NINJAL, 267-284.

Srdanović, I., Suchomel, V., Ogiso, T., Kilgarriff, A. (2013) Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen, *Proceeding of the 3rd Japanese corpus linguistics workshop*, Department of Corpus Studies/Center for Corpus Development, NINJAL, 229-238.

Srdanović, I. (2014). Corpus based collocation research targeted at Japanese language learners. *Acta linguistica asiatica.* 4 (2), 25-35.

Stefanowitsch, A. and Gries, S. T. (2003) Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8 (2), 209-243.

Suzuki, S. (1972) *Japanese Grammar and Morphology*. Mugi shobo.

Tsujimura, N. (1996) *An Introduction to Japanese Linguistics*. Blackwell Publishers.

Internet resources

Sketch Engine: http://www.sketchengine.co.uk/ (17.8.2019.)

Chunagon: https://chunagon.ninjal.ac.jp/bccwj-nt/search (10.10.2019.)

要旨 (Abstract in Japanese)

「日本語のコーパスにおけるイ形容詞–分布、パターン、語彙制約—」

イレーナ・スルダノヴィッチ
（ユライドブリラ大学プーラ）

本稿では、コーパス言語学の実証的手法および最先端の言語資源と語彙プロファイリングツールを用いて日本語の形容詞の使用について検討する。まず、分析に利用したリソースを紹介し、その重要性と特徴について述べる。次に、現代日本語大規模コーパスにおける形容詞の分布を分析することにより、使用頻度の高い形容詞、さらには複合形容詞を形成する生産性の高い形容詞および接尾辞の用法の多様性を明らかにする。続いて、形容詞が持つ3つの主要な役割のパターンの分布を分析し、形容詞によって役割とパターンに異なる使用傾向が見られることを示す。特に、形容詞の連体修飾用法に焦点を当て、用法のパターンの複雑さを明らかにし、形容詞の連体修飾用法のタイプをより詳細に分類する必要があることを指摘する。さらに、形容詞の連体修飾用法における語彙的制約を調査した結果、連体修飾用法を全く持たない、あるいはほぼ持たない形容詞を見出すことができた。