JAPANESE WORD SKETCHES: ADVANTAGES AND PROBLEMS

Irena SRDANOVIĆ
University of Ljubljana, SI
irena.srdanovic@gmail.com

Naomi IDA
Meiji University, JP
idanaomi2002@yahoo.co.jp

Chikako SHIGEMORI BUČAR
University of Ljubljana, SI
chikako.bucar@guest.arnes.si

Adam KILGARRIFF
Lexical Computing Ltd., UK
adam@lexmasterclass.com

Vojtěch KOVÁŘ
Masaryk University, CZ
xkovar3@fi.muni.cz

Abstract

In this paper, we present results of an evaluation of Japanese word sketches and address in detail issues that were observed by the evaluators. A word sketch presents a list of salient collocates of a word, organized by the grammatical relations holding between the word and its collocate. The word sketch functionality is incorporated into the Sketch Engine corpus query system and has been created for more than twenty languages so far, including Japanese. The issues that have been discovered in the evaluation of word sketches in Japanese are to be addressed for further enhancement of the word sketch functionality. Other tools and resources which are combined for use and influence the performance of the word sketches should also be looked over. We divide the issues into the following: 1) the lemmatizer and tagger in use, 2) the sketch grammar that is specifically written for Japanese, and 3) the corpus and statistical methods.

Keywords

word sketches, Japanese collocations, evaluation, corpus, language technologies

Izvleček

V prispevku predstavljamo rezultate ocenjevanja japonskih besednih skic in podrobno prikazujemo probleme in težave, ki smo jih opazili ocenjevalci. Besedna skica je seznam izstopajočih kolokacij neke besede, ki ga organizirajo slovnične relacije med besedo in drugimi besedami, ki skupaj sestavaljajo kolokacije. Funkcije besedne skice so vgrajene v korpusno orodje Sketch Engine in na voljo trenutno že v več kot dvajsetih jezikih, med njimi tudi v japonščini. Problemi in težave, ki smo jih odkrili med ocenjevanjem besednih skic v japonščini, moramo dalje proučiti za okrepitev funkcij besednih skic. Problemi in težave so pri naslednjih: 1) pri sistemu ugotavljanja osnovne oblike besede in označevalcu besednih vrst v rabi; 2) v slovnici za skice, ki je napisana posebej za japonščino; 3) pri korpusu in statističnih metodah.

Ključne besede

besedne skice, kolokacije v japonščini, evalvacija/ocenjevanje, korpus, jezikovne tehnologije

Acta Linguistica Asiatica, Vol. 1, No. 2, 2011. ISSN: 2232-3317
http://revije.ff.uni-lj.si/ala/index

1. Introduction

The word sketches automatically summarize a list of salient collocates of a word, organised by the grammatical relations holding between the word and its collocate. By *collocate*, we refer to the word that joins with the headword to form a collocation. For any headword, a list of its collocates is a list of the words that it combines with to give its collocations (Kilgarriff et al 2010). The word sketches were first prepared for the English language and used for the compilation of the Macmillan English Dictionary for Advanced Learners (Rundell 2002). Later on they were integrated into the Sketch Engine corpus query tool (Kilgarriff et al 2004), created for numerous languages, and used on a large scale for lexicography by a number of publishers. The word sketches for Japanese were first prepared in 2008, employing 400 million-word web corpus that is tokenised and POS-tagged using the ChaSen toolset[1] and English translation of POS tags. The word skech grammar written for Japanese covers more then 50 collocational and grammatical relations for the Japanese nouns, verbs, adjectives and adverbs (Srdanović et al 2008a).

The first formal quantitative evaluation of word sketches is performed for four languages, Dutch, English, Japanese and Slovene, as part of the Sketch-Eval mini-project. The evaluation is undertaken from a user perspective, with the main question being "is the collocation suitable for inclusion in a published collocation dictionary". The background of the evaluation method and the results for all four languages is described in Kilgarriff et al (2010). In this paper, we concentrate on results of the Japanese word sketches evaluation and discuss the issues discovered by the evaluators.

2. Japanese word sketches

The creation of Japanese word sketches required preparation of the following components:[2]

- A corpus.
  At the time of creation of the Japanese word sketches, there was no publicly available corpus that could be used inside the SkE tool. Therefore, a large-scale Japanese language web corpus, named JpWaC, was created. Its size is 7.3GB or 400 million words.
- Language processing tools used for processing the corpus data: tokeniser, lemmatiser and part-of-speech (POS) tagger.
  The morphological analyzer and part-of-speed tagger ChaSen was used for processing the JpWac data. The ChaSen tagset is quite detailed, with 88 tags,

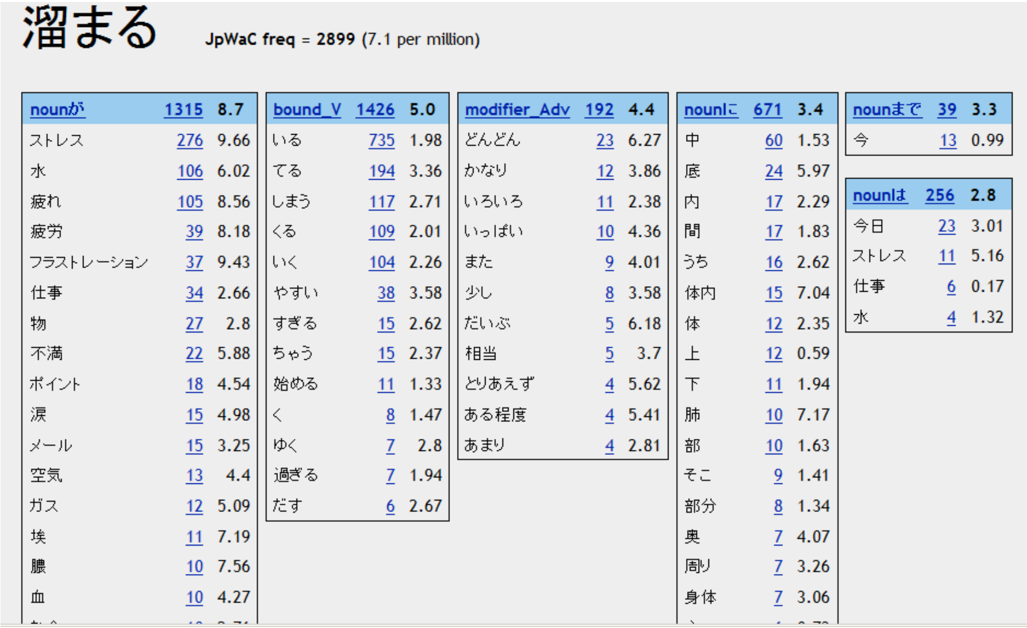
[1] http://chasen.naist.jp

[2] For details on the creation of the Japanese web corpus JpWaC and the Japanese word sketches, refer to Srdanović et al. (2008a).

and uses a fairly “narrow”, or fine-grained, tokenization: it splits inflectional morphemes from their stems. It uses the IPADIC dictionary.

- A sketch grammar[3] with a specified POS tagset for the language.
  The Japanese sketch grammar uses English translations of the ChaSen POS tagset. The Japanese sketch grammar defines 22 grammatical patterns covering more than 50 collocational relations for nouns, adjectives, verbs and adverbs.

The corpus was prepared in Sketch Engine format and installed in the system. The sketch grammar is also loaded into the system. Using the components described above, the system automatically selects frequent and statistically salient collocations. The statistics are based on the Dice coefficient.[4]



溜まる JpWaC freq = 2899 (7.1 per million)

nounが	1315	8.7
ストレス	276	9.66
水	106	6.02
疲れ	105	8.56
疲労	39	8.18
フラストレーション	37	9.43
仕事	34	2.66
物	27	2.8
不満	22	5.88
ポイント	18	4.54
涙	15	4.98
メール	15	3.25
空気	13	4.4
ガス	12	5.09
埃	11	7.19
膿	10	7.56
血	10	4.27

bound_V	1426	5.0
いる	735	1.98
てる	194	3.36
しまう	117	2.71
くる	109	2.01
いく	104	2.26
やすい	38	3.58
すぎる	15	2.62
ちゃう	15	2.37
始める	11	1.33
く	8	1.47
ゆく	7	2.8
過ぎる	7	1.94
だす	6	2.67

modifier_Adv	192	4.4
どんどん	23	6.27
かなり	12	3.86
いろいろ	11	2.38
いっぱい	10	4.36
また	9	4.01
少し	8	3.58
だいぶ	5	6.18
相当	5	3.7
とりあえず	4	5.62
ある程度	4	5.41
あまり	4	2.81

nounに	671	3.4
中	60	1.53
底	24	5.97
内	17	2.29
間	17	1.83
うち	16	2.62
体内	15	7.04
体	12	2.35
上	12	0.59
下	11	1.94
肺	10	7.17
部	10	1.63
そこ	9	1.41
部分	8	1.34
奥	7	4.07
周り	7	3.26
身体	7	3.06

nounまで	39	3.3
今	13	0.99

nounは	256	2.8
今日	23	3.01
ストレス	11	5.16
仕事	6	0.17
水	4	1.32

Figure 1 Japanese word sketch example for the verb tamaru (溜まる “to accumulate [intr.]”), partial results

[3] The sketch grammar is a mini-grammar of syntactic patterns. It is based on regular expressions over POS tags and enables the system to automatically identify possible collocations. See *Corpus Querying and Grammar Writing* on the Sketch Engine website, http://www.sketchengine.co.uk

[4] Refer to *Statistics used in the Sketch Engine* on the Sketch Engine website, http://www.sketchengine.co.uk.

Table 1 Types of collocational relations for verbs in the Japanese word sketches (14 different types of relations)

POS	Grammar sketch pattern	Type of relation	Example	Example transcription
Verb (14)	modifier_Adv	Adv modifying V	にこにこ笑う	*nikoniko warau*
	noun は	noun_*wa*＋V	彼は笑う	*kare wa warau*
	noun が	noun_*ga*＋V	鬼が笑う	*oni ga warau*
	bound_V	bound verbs connecting to free verbs	わらっちゃう	*warattyau*
	V_bound	free verbs connected to bound verbs	連れて行く	*turete iku*
	noun で	noun_*de*＋V	鼻で笑う	*hana de warau*
	noun に	noun_*ni*＋V	最後に笑う	*saigo ni warau*
	noun から	noun_*kara*＋V	（心の）底から笑う	*(kokoro no) soko kara warau*
	noun まで	noun_*made*＋V	最後まで笑う	*saigo made warau*
	noun を	noun_*wo*＋V	（人の）失敗を笑う	*(hito no) sippai wo warau*
	noun へ	noun_*he*＋V	公園へ行く	*kooen e iku*
	Coord	coordinate relation	笑う・泣く	*warau - naku*
	Suffix	V+*suffix*	笑いっぱなし	*waraippanasi*
	Prefix	*prefix*＋V	超笑う	*tyoo warau*

Figure 1 gives a word sketch for the Japanese verb *tamaru* (溜まる“to accumulate [intr.]”). Different grammatical relations, such as noun-particle-verb collocates (for example, noun が for noun+*ga*+verb collocates), bound verbs that appear with the verb, adverbs modifying the verb etc., are displayed in order of their significance, revealing the most frequent and most salient sets of collocations (see the first and second columns with numbers of frequency and saliency respectively).

Table 1 shows what types of collocational relations are covered in the Japanese word sketches for words classified as verbs (all together fourteen). An example for each of the relations is given.

3. Evaluation: method and results

Since the creation of the Japanese word sketches we have undertaken various kinds of evaluation. Srdanović et al (2008a) describes a comparison of a newspaper corpus and the JpWac corpus, showing that newspaper data are more specific both in terms of form (being written mainly in the past tense and not using the formal predicate form *masu/desu*) as well as content, with a high proportion of news-specific and politically-oriented nouns. On the other hand, JpWaC contains more informal and interactional material, and more diverse content. Later on, the study by Srdanović et al (2008b) explored the appearance of adverb-final modality forms in various corpora, confirming that the newspaper corpus, as well as some other corpora, is more specific in nature than the web corpus. This study shows that the web corpus is the most similar to the Balanced Corpus of Contemporary Written Japanese (Maekawa et al 2010). Srdanović & Nishina (2008) evaluate the functionality by comparing its results to the first and only collocation dictionary for Japanese language students (Himeno 2004). The comparison of randomly selected items in the dictionary with the word sketch for the same word clearly shows the much wider spectrum of collocational and grammatical relations as well as a richer variety of collocations in the sketches. We see great potential for using word sketches for future dictionary compilations

In this section we present the evaluation methods and results for Japanese word sketches in the Sketch-Eval project (Kilgarriff et al. 2010).

3.1 Type of evaluation

Sketch-Eval is a quantitative type of evaluation, undertaken from a user perspective. It measures precision, which is the percentage of the answers given that are correct.[5] It is measured by examining the word sketch responses with the critical question being “is the collocation suitable for inclusion in a published collocation dictionary”. Here, the Oxford Collocations Dictionary (OCD 2009) is proposed as a model and a reference point for what we wish to produce automatically. A number of human experts evaluated a sample of dictionary entries and a set of their collocates in the word sketch. For each language, forty-two headwords were sampled and twenty most salient collocates[6] for each of the headwords were inspected. Four languages were included in the Skech-Eval, among which was also Japanese.

[5] The information sciences distinguish between evaluating precision and recall. Kilgarriff et al (2010) describe it as follows: “Precision is the percentage of the answers given that are correct. Recall is the percentage of all correct answers that are found. If my word sketch for *flour* contains only *sift* and *sieve*, it has 100% precision, since all the given collocates are correct, but low recall, since there are many other collocates it does not give. As a response gets bigger, precision usually falls off (since some incorrect answers creep in) but recall improves (as more of the correct answers are included). Changing the size of the answer is a matter of adjusting the ‘precision/recall tradeoff."

[6] This was subject to the constraint that no more than two thirds relate to any single grammatical relation, to provide variety of collocational relations. Later on it was realized that it might have been better to have lower number of collocates for medium and low-frequency words. (Kilgarriff et al. 2010)

Table 2 Randomly extracted sample list and reserve list of words for evaluation of the Japanese word sketches

	Sample list			Reserve list		
	Nouns	Verbs	Adjectives	Nouns	Verbs	Adjectives
Common (top 2999)	急 研究 完成 男性 緑 評価	生まれる 扱う 支払う 忘れる	よろしい っぽい 素晴らしい 大きい	心配 箱 積極 プロセス 地区 建設	ちゃう 知れる 語る 受け入れる	重い 長い 忙しい 無い
Mid (3000-9999)	欠席 蓄積 マスター 俳句 情勢 有力	まつ づく 資する 溜まる	黒い おとなしい こい 親しい	フォント 刑事 澄 蝶 包装 メス	溢れる 拒む 隠る しむ	柔らかい むずかしい きつい とんでもない
Low (10,000-30,000)	クレイ 方角 近鉄 走り 苑 人妻	駆け込む やせる 書き留める 滅ぶ	むつかしい 仲良い くすい 腹立たしい	水槽 グローバリゼーション 青島 懇話 射精 モグラ	対する 振舞う 吸い上げる くぼむ	若々しい 面倒くさい 耐え難い 香ばしい

We took a sample from the 30,000 commonest nouns, verbs and adjectives in the corpus, in a ration of roughly 2:1:1, and with the sample structured as in Table 2. Within these constraints, the sampling was random. Table 2 shows our automatically extracted random sample (and reserve) list of Japanese words for the evaluation of the Japanese word sketches. The reserve list is also available to provide a replacement for a misanalyzed word in the sample list. In case of Japanese, the reserve list is also used instead of narrowly analyzed morphemes, such as adjectival suffix っぽい, *-ppoi*, "-ish, like" or for words that may be typically written in another orthography, such as まつ, *matu* (typically written as 待つ "to wait").

For the Japanese SkE evaluation, it was specific that noun-verb-adjective proportion in the selected word list was quite different from other languages. This is because the Japanese tagset ChaSen includes under the noun tag a) nouns being part of so called "*suru* verbs" (*suru*-V), formed as noun + verb "*suru*",[7] and b) adjectives ending in –*na*, derived from nouns.[8]

3.2 Evaluation categories and evaluators

A customised version of the Sketch Engine was prepared in which word sketches contained only the twenty highest-scoring collocates for each word, and in which each collocate was associated with a menu with the following items:

- Good
- Good but wrong grammatical relation
- Maybe (e.g. not striking collocate)
- Maybe (specialised vocabulary)
- Bad

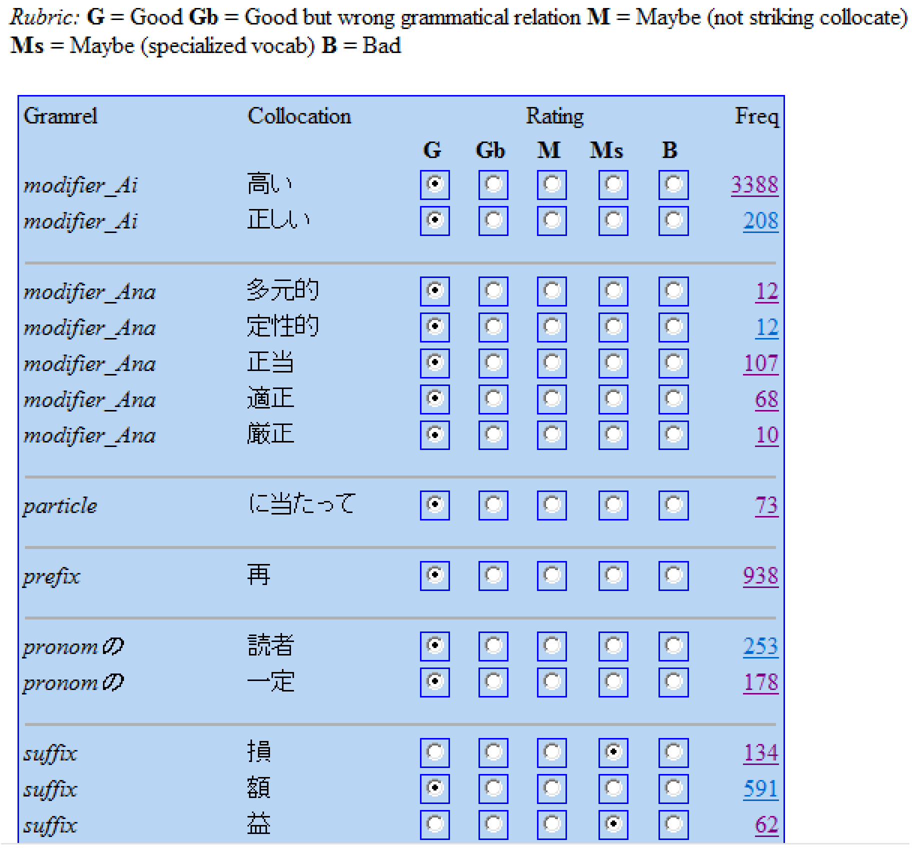
Three evaluators performed the Japanese SkE evaluation: two of them being native speakers of Japanese, language teachers and linguists, and the third one being a non-native speaker, language teacher, linguist and lexicographer.

A screenshot of the evaluators' word sketch interface is shown in Figure 2. Evaluators selected the relevant item from the menu and choices were stored in a database.

In order to rule out 'unclear' data, we distinguished those instances where all evaluators agree from those where they disagree, and based our results only on the agreement cases. We noted that agreement on the boolean decision, "good or not good" was substantially higher than agreement on finer-grained categories, so we merged "Good" and "Good-but" as "good" and all other categories as "bad".

[7] *Suru* is a verb which means "to do" in general and is highly productive to derive verbs of foreign origin, e.g. *kekkon* "marriage (noun)" vs. *kekkon suru*"to get married (verb)", *kopii* "copy (noun)" vs. *kopii suru* "to copy (verb)".

[8] There are two types of Japanese adjectives: adjectives that end in -*i*, e.g. *ookii* "big", and adjectives that end in –*na*, e.g. *genki-na* "healthy, vigor, good", derived from the noun *genki* "health".



Rubric: **G** = Good **Gb** = Good but wrong grammatical relation **M** = Maybe (not striking collocate)
Ms = Maybe (specialized vocab) **B** = Bad

Gramrel	Collocation	Rating					Freq
		G	**Gb**	**M**	**Ms**	**B**	
modifier_Ai	高い	●	○	○	○	○	3388
modifier_Ai	正しい	●	○	○	○	○	208
modifier_Ana	多元的	●	○	○	○	○	12
modifier_Ana	定性的	●	○	○	○	○	12
modifier_Ana	正当	●	○	○	○	○	107
modifier_Ana	適正	●	○	○	○	○	68
modifier_Ana	厳正	●	○	○	○	○	10
particle	に当たって	●	○	○	○	○	73
prefix	再	●	○	○	○	○	938
pronom の	読者	●	○	○	○	○	253
pronom の	一定	●	○	○	○	○	178
suffix	損	○	○	○	●	○	134
suffix	額	●	○	○	○	○	591
suffix	益	○	○	○	●	○	62

Figure 2 An example of evaluators' word sketch interface for the word *hyooka* (評価) "evaluation, assessment"

3.3 Results

Table 3 shows the total number of collocations evaluated by all assessors, the number for which all evaluators agreed, and for these, the number that were good and the number that were bad (where "bad" includes "maybe"). The total number of collocations evaluated by all assessors is slightly less than the maximum possible of 20 collocates for each of 42 headwords, owing to a different evaluator's choice on replacement of a sample word with a word from the reserve list, and from a range of minor omissions. For Japanese there was three-way agreement among lexicographers for less than half of the data, so we also give figures for two-out-of-three agreement, which provides more reliable evaluation results. Two-out-of-three agreement refers to the number of cases when any of two out of three evaluators agreed.

Table 3 Japanese word sketch evaluation results

Agreement	Total colls assessed	Evaluators all agreed on	**Good**	**Bad**	**% good**
Three-way agreement	747	294	278	16	94.56%
Two-out-of-three agreement		690	600	90	86.95%

In total 747 collocations were judged by all evaluators. All three evaluators agreed on 294 instances, and any of two out of three evaluators agreed on 690 instances. Of those, all three evaluators agreed that 278 collocates were "good" and 16 "bad", which results in 94.56% of collocates being good. Calculating the results that any of two out of three evaluators agreed, shows that 600 collocates were evaluated as "good" and 90 collocates as "bad", which gives result of 86.95% of collocates being good.

On examining possible reasons for the high proportion of disagreement between evaluators, we discovered that one out of three evaluators had a noticeably different approach towards collocations that are basically good, but not complete as a semantic or syntactic unit. If an evaluator uses concordance examples to check instances of a collocation candidate in the corpus, they will see the full collocation, which appeared as incomplete in the word sketch, and from that point it can be regarded as good and useful for a dictionary editor. In the evaluation process, for this type of collocations, a new selecting choice such as "Good but not complete" would be preferred. On the other hand, "Good but wrong grammatical relation" could be excluded as a selection choice for Japanese word sketches. Other possible reasons for disagreements in evaluation are related to treatment of collocations with technical terms in specialized fields, polysemic keywords, orthography issues etc. Some of the issues with examples are presented and discussed in detail in the following section.

Comparing Japanese word sketch evaluation results to other three languages, Slovene (71.1% good), English (70.7% good), Dutch (66.3% good) (Kilgarriff et al 2010), we can say that Japanese word sketches showed the best performance. For the other three languages too, two thirds or more of the collocations on which the assessors agreed were of publishable quality, which confims the word sketches being a valuable resource for lexicographic work in any of four languages.

4. Discussion

The quality of the word sketches functionality depends on the quality of its components: the corpus, POS-tagger, sketch grammar etc, which are described in section 2. Although the evaluation results for Japanese word sketches show a high precentage of good collocations, there are some issues that were discovered during the evaluation. Japanese word sketches have problems especially with some POS-tagger and morphological analyzer issues, some overlooked grammar rules/relations, corpus junk and problems on the level of orthography. This section addresses these issues in detail.

4.1 Issues with the tagger and morphological analysis

ChaSen’s analysis and tagset are very fine-grained which causes the following common problems: the excessive analysis of *suru* verbs (*suru*-V) and adjective on *–na* (*na*-Adj) (4.1.1), excessive analysis of derived and compound nouns (4.1.2), and problems related to the Japanese writing system (4.1.3).

4.1.1 Excessive analysis of *suru* verbs (*suru*-V) and adjective on *–na* (*na*-Adj)

The Japanese tagset ChaSen analyses *suru*-V into its most basic form that is tagged as noun, N.Vs [名詞-サ変接続] and the verb form *suru*, which is tagged as V.free [動詞-自] and further on analyzes into various segments depending on its inflectional forms. Similarly, the tagset analyses *na*-Adj into the basic noun form, N.Ana [名詞-形容動詞語幹], and the derivational suffix -*na*. The tagset does not distinguish if the tagged noun appears only as a noun or it is a part of its derived form of *suru*-V or *na*-Adj. This issue is already obvious from the automatically extracted sample list of nouns, verbs and adjectives (see Table 2), where adjectives ending in -*na* and *suru*-V were included under the noun POS category. This kind of analysis brings some consequences in the word sketch search and results. For example, it is not possible to search for word sketch results of *suru*-V or *na*-Adj independently from their basic noun category, which hides the real frequency of each of these two types of words, the noun and its *suru*-V pair, or the noun and its *na*-Adj pair. Also, the word sketch results provide lists of collocations that might be real collocates of only one of the word types. However, the issue with the real collocates is related to the current version of the sketch grammar for Japanese and could be overcome to some extent with corrections in the grammar (see also 4.2.3)

The consequences become more noticeable especially in the case of *suru*-V, when the meaning of the noun is quite different to that of the *suru*-V that is derived from the noun (1), (2).

(1)	N	マスター	*masutaa*	“teacher, leader”
	suru-V	マスターする	*masutaa-suru*	“to master”
(2)	N	左右	*sayuu*	“left and right”
	suru-V	左右する	*sayuu-suru*	“to control, affect, influence”

In the examples above, the meanings of the noun and that of the verb are very different but the word sketch search is possible only for the form マスター, *masutaa*, or 左右, *sayuu*, whether or not it is only a noun, or a derived verb form. Besides some corrections in the sketch grammar file, the issue is possible to resolve with an additional retagging, as shown in Table 4. The table suggests that *suru*-V and *na*-Adj are tagged with one separate tag, and thus differentiated from the nouns that never appear with *suru* or *-na*.

Table 4 Current and desired tagging of *suru*-V and *na*-Adj

POS (example)	Current tagging	Desired tagging
suru-V (*eigo wo masutaa suru* “to master English”)	*masutaa*[N.Vs] *suru*[V.*suru*]	*masutaa_suru*[*suru*-V]
noun, without possible *suru* form (*masutaa ga kita* “the teacher has come”)	*masutaa*[N.Vs]	*masutaa*[N.Vs]
na-Adj (*genkina ko* “healthy child”)	*genki*[N.Adj]+*na*	*genkina*[*na*-Adj]
noun, without possible *-na* form (*genki ga nai* “not to be well”)	*genki*[N.Adj]	*genki*[N.Adj]

4.1.2 Excessive analysis of derived and compound nouns

ChaSen divides derived and compound nouns into morphemes and treat them as different segments, separating their non-standalone prefixes and suffixes too. This kind of narrow analysis causes incomplete and misleading collocational results, as shown in the following examples.

(3) 優秀な 研究 | 者[9]
yuusyuu-na *kenkyuu* | *sya*
excellent research | -er
“excellent researcher”

The word sketch shows the noun *kenkyuu* “research” collocating with the noun *yuusyuu* “excellent”[10]. This is incomplete collocational information: it is the noun *kenkyuu-sya* “researcher”, derived from the noun *kenkyuu* “research”, that collocates with *yuusyuu-na.*

(4) 財界 の 有力 | 者
zaikai *no* *yuuryoku* | *sya*
financial circles possessive particle influential | person
“an influential person in financial circles”

[9] Inappropriate separation of derived or compound noun is denoted by vertical bar (|).

[10] The *na*-Adj *yuusyuu-na* is tagged as N, in the form of *yuusyuu*, as we mentioned above.

Similarly, the word sketch shows the word *yuuryoku* "influential"[11] collocating with the noun *zaikai* "financial circles". In reality, it is *yuuryoku-sya* "influential person" that collocates with *zaikai*.

(5) スポーツ | 用品 を 扱う
supootu | *yoohin* *wo* *atukau*
sport | goods object particle deal with
"to deal with sport goods"

The POS tagger divides compound nouns into segments, as shown in the example (5), which can be regarded as semantically incomplete from the point of view of collocational relations. In word sketches, only parts of compound nouns are shown as collocates. For example, the word sketch for *atukau* "to deal with" shows that the verb collocates with the noun *yoohin* "goods", which can be regarded as incomplete. The complete collocation is *supootu yoohin* "sport goods".

Similarly, there are examples of suffixes that are tagged as nouns (歳 *sai*, 代 *dai* etc.) and appear as separate words in the word sketches:

(6) 50 | 歳 の 男性
50 | *sai* *no* *dansei*
fifty | age possessive particle man
"a fifty-year-old man"

(7) 50 | 代 の 男性
50 | *dai* *no* *dansei*
fifty | generation possessive particle man
"a man in his fifties"

(8) 使用 | 料 を 支払う
siyoo | *ryoo* *wo* *siharau*
use | fee object particle pay
"to pay rental fee"

-*sai* in (6), -*dai* in (7) and -*ryoo* in (8) are the nouns with suffix role and inevitably require another noun, respectively 50 and *siyoo* "use, rental" to be complete in their usage.

Although the narrow analysis causes some obvious problem in finding appropriate collocations, it also offers the possibility of exploring behavior of suffixes and prefixes in detail. This type of information has also been a subject of interest for dictionary makers, language learners and language specialists, for example see Vance (1991).

[11] The *na*-Adj *yuuryoku-na* is tagged as N, in the form of *yuuryoku*.

What is encouraging from the lexicographers' point of view is that the complete and correct collocational relations for the majority of the above examples can be easily found from the combination of word sketch and corpus examples, searchable from the word sketch interface. However, when someone looks only at the list of collocates, the results are misleading, which gives rise to instability in the evaluators' judgments. This matter should be tested in detail among different POS taggers and dictionaries for further enhancement of the functionality.

4.1.3 Orthography issues: words that can be written in *hiragana*, *katakana* or *kanji*

This issue is peculiar to the Japanese language. The Japanese writing system uses a combination of three sets of letters: Chinese characters (*kanji*) and Japanese syllabic alphabets (*kana*: *hiragana* and *katakana*). There is a fair amount of fluctuation and overlap in the use of characters and *kana* (Seeley 1991). This problem is not yet very well addressed in the current natural processing tools, including ChaSen, and this is reflected in the word sketches. On the one hand, since there are two or three different orthographies for the same word, such as *subarasii* (9), the information about collocates is dispersed among different orthography variations, and therefore some collocates are missed. On the other hand, when there are two or more different words that are identical when written in *kana*, such as *matu* (10), results for the three words are mixed together.

(9) *i*-Adj: *subarasii* "wonderful, splended" 素晴らしい，すばらしい，スバラシイ

(10) V or N: *matu* まつ "to wait/pine/end", 待つ "to wait", 松 "pine", 末 "end"

Since the creation of the Japanese word sketches, some research progress has been made with the development of UniDic, the new dictionary for morphological analysis. The dictionary can be used with ChaSen or MeCab. The research on its accuracy reveals slightly better results in its combination with MeCab than with ChaSen.[12] The dictionary also provides some improvements in dealing with the Japanese orthography and pronunciation issues by providing canonical forms, word forms, writing variants, speech variants and accent.[13]

We plan to use the new set of tools for the new version of Japanese word sketches, which is expected to improve the performance of the functionality. This will include also the usage of canonical orthographic forms for each Japanese lemma, which means that each lemma will have only one, its typical, orthographic form.

[12] http://www.tokuteicorpus.jp/dist/

[13] The UniDic contained 150,000 words (canonical forms) (July 2009). (http://www.tokuteicorpus.jp/dist/) .

4.1.4 Other tagger issues

There are some other tagger issues that were revealed during the evaluation, such as errors in morphological analysis or a different inflectional form of an actual collocate than of its lemma form displayed in the word sketch

For example, the tagger analyses the proper noun 京急蒲田 *Keikyuu kamata*, a train station name, into three different elements, and the word sketch wrongly presents 急 *kyuu* and 蒲田 *kamata* as collocations

Other examples of wrong morphological analysis are in case of sayings, idioms, such as 急がば回れ *isogaba maware* "slow and steady wins the race". The morpheme 急 *kyuu* "sudden, quick" is again regarded as a separate lemma and as a collocate of the potential form 回れる *mawareru* "to be able to go around"

As for different inflectional forms, we find cases such as (11), where the lemma of the collocate is a dictionary form, while the actual collocate is only in negation. Another example (12) is a variant of the adjective *yorosii*, an honorific variant of the word *yoi/ii* "good, nice", which is present in a highly frequent set phrase.

(11)	忘れる	いける	→	忘れ（ては）いけない，（て）いけない
	wasureru	*ikeru*	→	*wasure(tewa) ikenai, (te) ikenai*
	Forget	able to go	→	"You should not forget"

(12)	よろしい	お願い	→	よろしくお願いします
	Yorosii	*onegai*	→	*yorosiku onegai simasu*
	Good	request	→	"I would appreciate your favor"

4.2 Issues with the word sketch grammar

The evaluation of word sketch results revealed some issues in the word sketch grammar that are general for sketch grammar syntax and in use for all the languages (4.2.1), and some issues that could be improved by additions or changes in the grammar rules for Japanese (4.2.2, 4.2.3 and 4.2.4).

4.2.1 The reach of collocation: more than two collocates

In this section we list examples of collocates for which an additional collocational element in the phrase is lacking. This issue is related to the general limitation of sketch grammar syntax for identifying all parts of collocations of three or more words.

(13)	グローバリゼーション	が	もたらす
	guroobarizeesyon	*ga*	*motarasu*
	globalization	subject particle	bring
	"the globalization brings"		

(13')	グローバリゼーション	が	もたらす	影響
	guroobarizeesyon	*ga*	*motarasu*	*eikyoo*
	Globalization	subject particle	bring	influence
	“the influence that the globalization brings”			

The word sketch shows the noun *guroobarizeesyon* “globalization” collocating with the verb *motarasu* “to bring” (13). However, *guroobarizeesyon ga motarasu* “globalization brings”, without an object, is not a complete phrase. The noun *eikyoo* is the head of this noun-modifying clause and is related to two words, the noun *guroobarizeesyon* and the verb *motarasu*. In the web corpus, we find many occurrences of the example *guroobarizeesyon ga motarasu eikyoo* “the influence that the globalization brings”, which suggests that the collocational relation is fully established only when all of three elements are present

Another similar example is a collocational relation between nouns without the head noun.

(14)	グローバリゼーション	の	負	（の	側面）
	guroobarizeesyon	*no*	*hu*	*(no*	*sokumen)*
	globalization	poss. particle	negative	(poss. particle	side)
	“the negative side of globalization”				

According to word sketch results, *guroobarizeesyon* and *hu* are collocates (14). However, this can be regarded as incomplete and with no semantic and syntactic relation established, without the pivot noun *sokumen*.[14]

These kinds of examples cannot be easily judged as correct or bad collocates. Depending on an evaluator’s judgment, the semantic or syntactic relation of such collocates can be brought to question. However the word sketch results give a good hint to lexicographers to further check the collocates in corpus examples and search for the full collocational relation. Also, a method for identifying collocations longer than two words is currently being developed, for use in future evaluations.

[14] Similar example in "A Quantitative Evaluation of Word Sketches" (Kilgarriff et al. 2010):

“if the system lists a word which only collocates with the headword within a three-or- more-word unit, as *put* is a collocate for *cat* only in the context of *out* (“put the cat out”), is the collocate good or bad? Our decision was to treat it as good, as it is enough to signal to say to a lexicographer that there is a collocation to be included in a collocation dictionary, even if the system has not found all of it. But it was not a decision that human evaluators were comfortable with.”

“Multi-word items were a recurring concern, as it did not seem natural to the evaluators to mark *cat* as good at *put* when the word sketch gave no indication that *out* was also needed: this was the evaluators' most often –voiced concern.”

The example above is a problem of the grammatical level, while the problem of (15) in this paper is a problem of the lexico-syntactic level.

4.2.2 Distance of collocates

(15)	日本総領事館	に	亡命	を	求めて	駆け込んだ
	nihon-sooryoozikan	*ni*	*boomei*	*wo*	*motomete*	*kakekonda*
	the Japanese Consulate General	to	exile	obj.part.	want	rushed

“He rushed to the Japanese Consulate General for exile”

(16)	最寄り	の	交番	に	息	を	切らして	駆け込む
	moyori	*no*	*kooban*	*ni*	*iki*	*wo*	*kirasite*	*kakekomu*
	nearest	of	police box	to	breath	object particle	grasp	to rush

“to rush into the nearest police box grasping for breath”

The word sketch shows the verb *kakekomu* “to rush into [the past tense form *kakekonda*]” collocating with *boomei* “exile” in the example (15), *iki* “breath” in the example (16). However, they are not direct collocates since they are syntactically positioned at a distance, in distinct clauses (here an inserted clause functioning as an adverbial clause). In both cases above, another verb with a predicate function exist as a direct collocate of the nouns in question, that is *motomeru* in *boomei wo motomete* “to seek for an exile” and *kirasu* in *iki wo kirasite* “to grasp for a breath”. The issue is closely related to the problem of too wide grammatical relations (4.2.4) and the sketch grammar should be corrected for the grammatical relation in question to include only the first predicate verb as a collocate and exclude elements from another clause and thus limit the syntactic range (reach) where collocation extends to. However, the possibility to reach so distant collocates with the sketch grammar can be a valuable source for lexical information. This kind of distant co-occurrence of words could be added to the sketch grammar as a different type of collocational relation, called, for example, “distant collocates”.

4.2.3 Missing collocates / sketch grammar relations

As mentioned above, the current sketch grammar for Japanese has a set of 22 collocational patterns, which covers various collocational relations for nouns, verbs, adjectives and adverbs: 16 different types of collocational relations for nouns, 14 for verbs, 7 for adjectives ending in *–i*, 11 for adjectives ending in *-na*, and 1 for adverbs. However, Sketch-Eval and other word sketch evaluations have descovered that a few sketch grammar relations are missing in the current sketch grammar.

The most important one from the point of view of overall lexical coverage is the inclusion of *suru*-V in various collocational relations for verbs. As mentioned in section 4.1, this issue is related to the tagging of a part of *suru*-V as nouns, which is currently overlooked in some sketch grammar relations. The correction of the grammar would resolve some of the issues.

For example, if we search for collocations of the noun *masutaa* “teacher, leader”, we get collocational relations for nouns only, where the noun *masutaa* is the keyword; we do not get collocational relations for verbs, where the noun part of the verb *masutaa suru* “to master” is the keyword. Thus, the word sketch provides collocational relations such as *masutaa wo yobu* “to call a teacher”, but does not provide collocations such as *tsukaikata wo masutaa suru* “to master the usage”, The same is true, if we exchange the keywords, for example, if someone searches for collocates of *tsukaikata* “usage”, *suru*-V, such as *masutaa suru* will not be in the collocate list of verbs. This deficiency can be overcome with corrections in the sketch grammar, but with the current tagset the collocational results for both. The noun and the *suru*-V would be still present without distinction in the same word sketch page.

Another missing point of the sketch grammar is for the collocational type noun_noun, or noun_noun_noun, which would cover compound noun collocations such as *Nihongo kyooiku* (日本語教育) “Japanese language teaching” or *Nihongo kyooiku gakkai* (日本語教育学会) “The society for Japanese language teaching”.

These types of deficiencies were not exposed in the Sketch-Eval type of evaluation since they relate to recall rather than precision. We plan to overcome the issues mentioned above in the new vesion of the sketch grammar for Japanese.

4.2.4 Too wide sketch grammar rules

The word sketch evaluators noticed that some collocations are bad or good but not so striking since sketch grammar rules are too wide for a few collocational relations. This is especially in the case of collocational relations for bound nouns and coordinating relations.

For example, one case of coordinating relations that is evaluated as poor is *subarasii* “great, superb” and *kazuooi* “many, numerous”, which actually is not a type of coordinate relation, as can be seen in the corpus example (17).

(17)	素晴らしい	映画	を	数多く	作り だす
	subarasii	*eiga*	*wo*	*kazuooku*	*tukuridasu*
	great	movie	object particle	many	create

“to create many great movies”

The sketch grammar rule for this kind of relations needs more contstraints for better results.

Too wide sketch grammar rules are related to the already described issue of collocates in different clauses, in other words, distant collocates (4.2.2).

4.3 Issues with the corpus and statistical methods

There are two main issues that appear in relation to the corpus and the statistical measurement:

- Page duplicates: when the same pages (or their copies) appear a number of times in the corpus. The result of this is that the same information appears a number of times in the corpus, and the word sketch results wrongly present that some collocations are frequent, which they are not.

- Salience related problems happen when some collocates appear very frequently but only from one source, which is from one web page in the case of web corpora. Therefore, the results would be more accurate if the current statistical measurement took into account that collocates that appeared multiple times from one source were less salient.

Examples of collocation candidates that result from page duplicates or only one source and that are marked as bad are: *kesseki ga rongai* (欠席が論外) "absense is out of the question", *wakawakasii midori* (若々しい緑) "youthful green", *Terayama no haiku* (寺山の俳句) "Terayama's haiku", *tikuseki to seitoo* (蓄積と正統) "accumulation and legitimacy".

In addition, evaluators indicated that they would welcome a genre classification of the corpus, which would make the word sketches more usable in the field of Japanese language education. This kind of classification could be well applied in lexicography and other fields too.

The Japanese web corpus that is currently in use was created four to five year ago. Therefore, a new up-to-date web corpus with more thorough check on page duplicates or on other possible junk information would be welcome.

5. Conclusion and further work

The evaluation of Japanese word sketches inside a mini-project Sketch-Eval, though with a small number of evaluators, proved to be helpful both for system developers and system users, and especially promising for further research and activities on both sides and in collaboration. The evaluation results show a high percentage of good collocations, and we can conclude that Japanese word sketches could be a very useful resource for creation of collocational dictionaries. However, there are some issues that were discovered during the evaluation and which call for further enhancement of the functionality and for improvement of various components used by the tool (morphological analyzer/POS tagger, corpus, sketch grammar).

One of the basic questions to confront was the range of "collocation" and what set of words can be regarded as a collocate. Since word sketches were able to reach quite far in the search for collocates, sometimes neglecting our sense of syntax and semantics in a very general sense, we were newly confronted with sets of words which surprised us but opened our eyes. The word sampling and word sketches processed for the test evaluation suggested that linguists and language teachers should not limit themselves to their intuition and limited way of logic, but should be constantly ready for unusual (and actually at times usual for non-native learners') points of view. Of course, we are also aware that there are some language-specific categories and parts of speech for which it is necessary to develop an umbrella category with subcategories in order to expect more effective results with word sketches.

A similar evaluation project with corrected POS tags, morphological analysis and evaluation categories, as well as increased number and variety of judges will probably show better results and offer yet new ideas for dictionaries, textbooks and teaching methods. Prior to this evaluation, a new version of Japanese word sketches will be created with a novel set of components, which would include a new and more up-to-date web corpus for Japanese, another morphological analyzer and an improved sketch grammar.

The issues related to POS taxonomy and Japanese orthography are actually problems which are very well present in the contemporary Japanese language grammar and orthography system. Word sketches reflect these problems "faithfully".

Aknowledgement

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536, in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project P401/10/0792.

References

Himeno, M. (2004). *Nihongo hyoogen katuyoo ziten*. Kenkyusha

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX*. France: Université de Bretagne. 105-116. (available at http://promethee.philo.ulg.ac.be/engdep1/download/bacIII/sketch-engine.pdf)

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I., Tiberius, C. (2010). A Quantitative Evaluation of Word Sketches. *Proceedings of the XIV Euralex International Congress.* Leeuwarden : Fryske Academy. 7pp. (available at http://nlp.fi.muni.cz/publications/kilgarriff_xkovar3_etal/kilgarriff_xkovar3_etal.pdf)

Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H., Den, Y. (2010). Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese. *Proceedings of LREC 2010*, Malta. 1483-1486.

Oxford Collocations Dictionary for Students of English (OCD). (2009). Oxford University Press

Rundell, M, ed. (2002). *Macmillan English Dictionary for Advanced Learners*. London: Macmillan.

Seeley, C. (1991). *A History of Writing in Japan*. University of Hawai'i Press, Honolulu. 243pp.

Srdanović, E. I., Erjavec T. & Kilgarriff, A. (2008a). A web corpus and word-sketches for Japanese. *Sizen gengo syori (Journal of Natural Language Processing*) 15/2. 137-159. (also available at http://www.jstage.jst.go.jp/article/imt/3/3/3_529/_article)

Srdanović, I, Bekeš, A., Nishina, K. (2008b). Distant collocations of adverbs and modality forms observed in various Japanese language corpora. *Tokutei ryooiki kenkyuu 'Nihongo* *koopasu', Tokyo: Monbukagakusyoo kagakukenkyuuhi tokuteiryooiki kenkyuu 'Nihongo koopasu' Sookatu ban* (*Workshop of the Priority Area Research "Japanese corpus"),* Tokio. 223-230.

Srdanović, E.I., Nishina, K. (2008). Koopasu kensaku tuuru Sketch Engine no nihongoban to sono riyoo hoohoo (The Sketch Engine corpus query tool for Japanese and its possible applications), *Nihongo kagaku* (*Japanese Linguistics*) 23. 59-80.

Vance, T. J. (1991). *Instant vocabulary through prefixes and suffixes.* Power Japanese series. Kodansha International. 128pp.