

# A Web Corpus and Word Sketches for Japanese

Irena Srdanović Erjavec<sup>†</sup>, Tomaž Erjavec<sup>††</sup> and Adam Kilgarriff<sup>†††</sup>

Of all the major world languages, Japanese is lagging behind in terms of publicly accessible and searchable corpora. In this paper we describe the development of JpWaC (Japanese Web as Corpus), a large corpus of 400 million words of Japanese web text, and its encoding for the Sketch Engine. The Sketch Engine is a web-based corpus query tool that supports fast concordancing, grammatical processing, ‘word sketching’ (one-page summaries of a word’s grammatical and collocational behaviour), a distributional thesaurus, and robot use. We describe the steps taken to gather and process the corpus and to establish its validity, in terms of the kinds of language it contains. We then describe the development of a shallow grammar for Japanese to enable word sketching. We believe that the Japanese web corpus as loaded into the Sketch Engine will be a useful resource for a wide number of Japanese researchers, learners, and NLP developers.

**Key Words:** *Japanese web corpus, Corpus query tool, Sketch Engine, Word sketches*

## 1 The Sketch Engine

Of all the major world languages, Japanese is lagging behind in terms of publicly accessible and searchable corpora. This paper reports on the development of JpWaC (Japanese Web as Corpus), a large corpus of Japanese web text, which has been loaded into the Sketch Engine, where it takes its place alongside Chinese, English, French, German, and a range of other languages.

The Sketch Engine<sup>1</sup>(Kilgarriff et al. 2004) is a corpus tool with several distinctive features. It is fast, giving immediate responses for most regular queries for corpora of up to two billion words. It is designed for use over the web. It works with all standard browsers, so users need no technical knowledge, and do not need to install any software on their machine. It has been used for dictionary compilation at, amongst others, Oxford University Press<sup>2</sup>, Macmillan (Kilgarriff and Rundell 2002), Chambers Harrap and Collins, and also for language teaching (e.g. Chen et al. 2007) and language technology development (e.g. Gatt and van Deemter 2006; Chantree et al. 2005).

As well as offering standard corpus query functions such as concordancing, sorting, filtering

---

<sup>†</sup> Tokyo Institute of Technology

<sup>††</sup> Jožef Stefan Institute

<sup>†††</sup> Lexical Computing Ltd. and Universities of Leeds and Sussex

<sup>1</sup> <http://www.sketchengine.co.uk>

<sup>2</sup> <http://www.askoxford.com/oec>

etc., the Sketch Engine is unique in integrating grammatical analysis, which makes it possible to produce *word sketches*, one-page summaries of a word's grammatical and collocational behaviour as illustrated in Figure 1.

Figure 1 gives a word sketch for the noun お湯 (*oyu*). Different grammatical relations, such as Ai and Ana adjectives modifying a noun, noun-particle-verb collocates (を verb, で verb, が verb), noun-pronominal relations (with の) etc., are displayed in order of their significance, revealing the most frequent and most salient sets of collocations (the first and second column with numbers respectively). The list of collocates nicely reveals the different senses of the noun: (1) 'hot/warm water (water that is heated)' (お湯を沸かす, 熱いお湯, ポットのお湯, 鍋のお湯, お湯で茹でる, お湯で温める), (2) 'bath (water for taking a bath)' (お湯に入る, お湯から上がる, 湯

### お湯 JpWaC freq = 2876

modifier Ai 186 8.1	を verb 1339 5.3	で verb 332 5.0	は Adj 56 2.8	pronom の 397 1.6
ぬるい 10 8.68	沸かす 154 11.04	溶く 15 9.18	ぬるい 7 8.43	湯舟 7 8.91
熱い 78 8.12	わかす 16 8.32	溶かす 15 7.7	熱い 8 4.86	焼酎 23 8.38
あったかい 5 7.36	注ぐ 112 8.11	茹でる 13 7.59	か Adj 73 2.5	やかん 7 8.32
温かい 15 6.76	ためる 15 6.91	洗い流す 6 7.32	ぬるい 5 7.9	浴槽 10 8.23
暖かい 8 5.27	溜める 9 6.51	温める 15 7.28	熱い 8 4.85	ポット 11 8.17
やさしい 6 4.8	汲む 8 6.32	薄める 8 7.15	いい 7 0.13	少量 8 7.69
いい 13 1.02	足す 11 5.85	ゆでる 6 6.86	に verb 332 2.5	シャワー 20 7.35
	温める 7 5.78	洗う 22 5.83	浸かる 50 9.37	風呂 40 6.65
	張る 35 5.76	煮る 5 5.3	つかる 26 9.16	め 41 6.34
	はる 18 5.72	割る 7 4.61	漬ける 9 8.69	鍋 5 5.2
	貯める 6 5.33	流す 13 3.71	溶かす 18 7.97	温泉 5 4.45
	沸く 6 5.23	飲む 7 1.6	つかう 20 6.52	用 10 1.72
	いれる 11 4.99	入れる 5 0.04	溶ける 6 5.28	度 10 1.69
	混ぜる 6 4.75	が verb 366 3.6	入れる 37 2.93	以上 7 1.66
	洗う 10 4.58	沸く 42 8.33	つける 24 2.48	分 5 0.74
	入れる 96 4.3	冷める 8 6.34	入る 18 1.06	ここ 6 0.34
	捨てる 17 4.12	あふれる 9 4.55	modifier Ana 34 1.6	particle 103 0.7
	かける 20 4.09	使える 7 2.86	透明 7 4.77	だけ 8 0.21
	かぶる 5 4.06	出る 96 2.75		も 62 0.17
	浴びる 6 4.05	でる 16 2.67	は verb 98 1.6	suffix 104 0.4
	飲む 19 3.02	流れる 5 2.07	沸かす 8 7.83	割り 70 8.98
	流す 6 2.55	入る 12 0.47	から verb 32 1.5	屋 5 1.39
	切る 8 2.43		上がる 11 2.66	coord 30 0.1
	食べる 18 2.1			水 11 2.78
	待つ 10 2.08			

Fig. 1 Word sketch for お湯 (*oyu* 'hot water'). The first number is the number of instances for the collocation. The second is the association score (based on the Dice coefficient) for the collocation, which is used for deciding which collocations to show and sorting.

舟のお湯, 浴槽のお湯, 風呂のお湯, 易しいお湯, いいお湯)<sup>3</sup>The “coord” relation reveals the coordinate noun of お湯, that is 水 (*mizu* ‘water’).

Based on the grammatical analysis, we also produce a *distributional thesaurus* for the language, in which words occurring in similar settings, sharing the same collocates, are put together (Sparck 1986; Grefenstette 1994; Lin 1998; Weeds and Weir 2005), and “sketch diffs”, which compare related words (see section 3 below). The Sketch Engine is accessible either by a person using a web browser, or by program, with the output from the Sketch Engine (in plain text, XML or JSON) being an input to some other NLP process.

## 2 The JpWaC corpus

In this section we review the steps taken to compile the JpWaC corpus, give a quantitative account of its contents and discuss the validity of this kind of corpus for general language research. The corpus has been gathered using methods as described by (Sharoff 2006a, 2006b; Baroni and Kilgarriff 2006), and for Japanese (Ueyama and Baroni 2005). The corpus consists of texts obtained from around 50,000 web pages and contains just over 400 million tokens.

### 2.1 Compiling the corpus

For easy web corpus compilation, a great step forward has been made by the publicly available BootCat<sup>4</sup>(Baroni and Bernardini 2004; Baroni et al. 2006), WAC<sup>5</sup>(Web as Corpus Toolkit) and related work of the Wacky Project<sup>6</sup>(Baroni and Bernardini 2006). These open source tools provide the functionality for compiling a basic web corpus, and are straightforward to install and use on a Linux platform.

The main steps for compiling JpWaC were the following, explained in more detail below:

- (1) obtain list of Japanese language URLs
- (2) download the Web pages
- (3) normalize character encodings
- (4) extract metadata
- (5) remove document boilerplate
- (6) annotate with linguistic information

<sup>3</sup> Checking word sketches for the variant 湯 (*yu*), reveals an additional sense, ‘hot spring’, that is not present in the お湯 (*oyu*) variant.

<sup>4</sup> <http://sslmit.unibo.it/~baroni/bootcat.html>

<sup>5</sup> <http://www.drni.de/wac-tk/>

<sup>6</sup> <http://wacky.sslmit.unibo.it/>

1. The list of URLs to build the corpus is critical for any web corpus, as it determines the overall composition, in terms of language(s), registers, lexis, etc. In order to obtain a good corpus of general language, the URL list for JpWaC list was obtained via the methodology described in (Sharoff 2006a). First, the top 500 non-function words of the British National Corpus (see <http://natcorp.ox.ac.uk>; hereafter BNC) were translated into Japanese<sup>7</sup>. Random 4-tuples were generated from this list and the Google API was used to query the internet with them. The set of retrieved pages gave the final URL list.

2. We downloaded the HTML pages with the WAC toolkit program `paraget`, which implements parallel downloading of pages, while taking care not to overload particular servers; it also separates the large number of retrieved files into separate tree-structured directories.

3. For normalizing the character sets of the downloaded pages we used the BootCat tool `encoding-sort.pl`, which implements a heuristic to guess the encoding of each HTML page, and then the standard UNIX utility `iconv` to convert the actual file to UTF-8.

4. The extraction of meta-data implemented in BootCat/WAC, in particular title and author, did not work well on Japanese pages, so the only metadata we retained is the URL of each document.

5. For boilerplate removal we used BootCat `clean_pages_from_file_list.pl`, which removes HTML tags, JavaScript code and text-poor portions of the pages, such as navigational frames, leaving only the pure text of each page.

6. Finally, we annotated the corpus using ChaSen<sup>8</sup>, which segments the text into sentences, tokenises it, lemmatises the words and assigns them morphosyntactic descriptions (MSDs). The MSDs that ChaSen returns are in Japanese: to make it easier to understand and use the annotations by non-Japanese speakers (and those without Japanese language keyboards), we translated the Japanese MSDs into English.

## 2.2 Copyright

There has been much discussion amongst corpus-builders about copyright. Many have taken the view that, for any text to be included, the copyright holder must be contacted and the document can only be included if their consent is granted.

This perspective dates back to a pre-web era. A corpus tool such as Sketch Engine, when loaded with a web corpus, is performing a task that is equivalent to a search engine such as Google or Yahoo. In each case, the owners of the service have automatically gathered large numbers of

---

<sup>7</sup> The words were translated by Moto Ueyama.

<sup>8</sup> <http://chasen.naist.jp/>

web pages, processed them, and indexed them. Then, users are shown small extracts.

This basic function of search engines has not been legally challenged. Google and Yahoo do not ask copyright owners for permission to index their pages. They, like us, observe the “no robots” convention for not trawling pages where the owner has specified in a “robots.txt” file that they should not.

Various related search engine functions have been challenged, such as where Google preserves, and shows users, cached copies of pages which are no longer freely available because the owner of the material charges for access to their archive. There have also been challenges where copyrighted videos have been placed on YouTube: YouTube offers the defence that it did not know that the material was there, which is an acceptable defence (in US law) provided that YouTube promptly removes the offending material when asked to do so. The current issue in this case (YouTube vs. Viacom, a large case which is ongoing as at October 2007) is how quickly and effectively YouTube responds, when the copyright owner brings it to YouTube’s attention that it is hosting copyrighted material. Taking note of such cases, Lexical Computing Ltd has mechanisms for promptly removing copyrighted material from its corpora should a copyright owner object.

There is no legal question over the legitimacy of the core activity of commercial search engines, or, thereby, of the legitimacy of presenting web corpora in the Sketch Engine.

### 2.3 Corpus statistics

In this section we give some statistics over the corpus in terms of overall size, the number of web sites and documents, and the statistics over linguistic categories.

The total corpus size is 7.3 GB, or 2 GB if compressed (.zip). An overview of the corpus size in terms of web documents is given in the left side of Table 1. The documents, just under 50 thousand, correspond to Web pages, each one identified by its URL, e.g. <http://www.arsvi.com/0e/ps01.htm>. The number of hosts (16,000) corresponds to distinct Web sites, such as <http://www.arsvi.com>. Dividing the former by the latter gives us the average number of pages per host, 3.1. The last two lines in the table give the numbers of pages from each of the two domains present in the corpus, with three quarters of the pages being from .jp and one quarter from .com.

It is also interesting to see what kinds of websites the documents come from. The center of Table 1 gives the keywords, defined as alphabetic strings appearing in URLs, which cover more than a thousand documents. The keywords to a certain extent reflect the Web sites with the greatest number of documents, in particular [blog.livedoor.jp](http://blog.livedoor.jp) (1,646 documents), [www.amazon.co.jp](http://www.amazon.co.jp) (1,048 documents), [d.hatena.ne.jp](http://d.hatena.ne.jp) (759 documents), and [blog.goo.ne.jp](http://blog.goo.ne.jp) (690 documents).

Linguistic processing with ChaSen gave us 12,759,201 sentences and 409,384,411 tokens; statis-

**Table 1** Corpus statistics

<i>n</i>	What	Docs	URL Keyword	Tokens/Doc	Stats
49,544	Docs	6,486	Blog	8,263	Average
16,072	Hosts	3,471	Nifty	5,001	Median
3.1	Docs/host	2,362	Archives	3	Min
34,911	URL .jp	2,075	Livedoor	60,929	Max
14,633	URL .com	1,545	Diary		
		1,428	News		
		1,380	Cocolog		
		1,296	Exblog		
		1,051	Amazon		
		1,013	Archive		
		1,006	Geocities		

tics over tokens per document is given on the right side of Table 1.

## 2.4 JpWaC validity, and comparison with a newspaper corpus

At this point, the reader will naturally want to know what sort of language there is in the web corpus. People are often concerned that web corpora will give a partial and distorted view of a language. These are difficult concerns to address because there is no simple way to describe a collection, which is, by design, heterogeneous, and where there has been no process of assigning labels like “journalism” or “novel”, to individual pages. This is a central problem in corpus linguistics (Kilgarriff 2001).

One straightforward strategy is to make comparisons with other corpora. As newspaper corpora have been widely used in Japanese language research and we had at our disposal the data from Mainichi Shimbun newspaper for the year 2002 (hereafter News), we compared JpWaC with that. The raw News corpus was processed with ChaSen, in the same way as JpWaC, and has around 30 million tokens. We analyzed the differences between the two corpora with frequency profiling, as described in (Rayson and Garside 2000). The method can be used to discover items of interest (say, key words or grammatical categories) in the corpora, which differentiate one corpus from another. In our case, we used it to discover lexical items (word form + ChaSen tag) or just tags in isolation. The method is straightforward: we first produce frequency lists and then calculate the log-likelihood statistic for each item. LL takes into account the two frequencies of the item and the sizes of the respective corpora and the larger it is, the more the item is salient for one of the two corpora. Starting with the items that have the highest LL scores, it is then possible to investigate the major differences between JpWaC and News.

The highest LL scores resulted from differences in coding practices. Both corpora use Unicode, which allows for coding the alphanumeric characters and basic punctuation in the standard ASCII range or, for Japanese, in the “Halfwidth Katakana” block (U+FF00 – U+FFEF). Similarly, the space character exists in ASCII, but also as the Unicode character “Ideographic space” U+3000. This space (analysed as a token with the tag Sym.w by ChaSen) is the item with the highest LL score. It appears over 24 times per 1000 words in News, while hardly ever in JpWaC. JpWaC uses, for the most part, ASCII characters for Western spaces, letters and numbers, while News uses their Eastern Unicode equivalents. Furthermore, ChaSen tokenises words written in ASCII into individual characters (and tags them with Sym.a). Being unaware of such differences can have practical consequences in cases where a researcher is interested in patterns including these characters.

Table 2 shows the twenty most salient log-likelihood differences in ChaSen tags between the two corpora. As explained, the first one marks Ideographic space, the second ASCII letters, and the third (Sym.g) ASCII punctuation. The fourth shows that unknown words (i.e. words that are not in the ChaSen dictionary) are much more frequent on the web than in the newspaper corpus. These three high LL scores thus stem for a combination of different coding practices, coupled with differences in ChaSen processing. The fifth row (N.Num) shows that numerals are more frequent in News. (This is a true difference in content, as ChaSen correctly tags both ASCII numbers and their Eastern Unicode equivalents with N.Num.) More frequent in News are also measure suffixes and proper nouns. Bounding nouns, pronouns, auxiliaries and adverbs are much more frequent in JpWaC.

**Table 2** Differences in ChaSen tags between the JpWaC and News (first 20 tags). The first column gives the ChaSen tag, the second the log-likelihood score, and the third and fourth the frequencies per thousand words

Tag	LL	JpWaC	News	Tag	LL	JpWaC	News
Sym.w	3943641	0.02	24.54	N.Suff.p	179445	0.76	3.97
Sym.a	3757462	67.38	0.71	N.Pron.g	167407	9.97	3.44
Sym.g	526014	32.52	11.53	Aux	156444	70.35	51.72
Unknown	413833	14.85	2.99	N.Vs	153786	53.00	70.82
N.Num	367879	49.24	76.65	P.Conj	122523	32.72	21.67
N.Suff.msr	308487	11.10	23.88	N.Prop.p.c	111389	3.57	7.96
N.Prop.p.g	238821	4.46	11.98	Adv.g	99256	8.8	3.92
N.Prop.o	216779	2.26	7.65	N.Suff.g	86295	16.97	24.65
N.bnd.g	206389	18.52	8.31	Sym.bo	85666	14.81	21.99
N.Prop.n.s	204396	3.62	9.91	Adn	85075	7.76	3.51

In JpWaC the following forms / grammatical categories are more salient than in News:

- auxiliary verbs forms **ます** , **です** , **まし** , **ませ** showing that sentence endings in the web data, in contrast to the newspaper data, frequently uses the masu/desu form
- forms expressing modality **か** , **でしょ** , **よう** , **思う** , **ので** , **わけ**
- forms expressing politeness **お** , **ご** , **ざい**
- forms expressing informal language **って** , **よ** , **ね** , **ん**
- person and place deixis: the first personal pronoun **私** , and place deixis **この** , **その** , **それ**
- noun bound forms **こと** and **もの**

In the newspaper corpus the following are more salient than in JpWaC:

- past tense form **た** showing that the newspaper data is mainly written in the past tense, in contrast to the web data.
- various suffixes and prefixes expressing time, place, numerals, measures, people **日** , **市** , **県** , **役** , **円** , **メートル** , **際** , **人** , **さん**
- adverbial nouns expressing time **昨年** , **後** , **元**
- proper nouns **東京** , **大阪** , **鈴木**
- general nouns that are specific for newspapers content and politically oriented **容疑** , **首相** , **米国** , **テロ** , **自民党** (No nouns that are specific to web data occurred amongst the 100 highest-LL entries. We believe this is because Web data is more heterogeneous, so, while there are very many nouns which have higher relative frequency in JpWaC than News, none had high enough frequency and bias to reach the 'top 100'.)

Table 3 gives the twenty words (together with their ChaSen tag) that have the highest LL scores. Based on the top hundred words, the analysis reveals a number of interesting differences between the corpora.

The comparison of the two corpora shows that newspaper data is more specific both in terms of form (being written mainly in past tense and not using masu/desu), as well as content (high proportion of news specific nouns). On the other hand, JpWaC contains more informal and interactional material, and more diverse content. The distinction between formal, distanced language and informal, interactional language is the most salient dimension of language variation, as has been explored and established in a number of studies by Biber and colleagues (Biber 1988, 1995), see also (Heylighen 2002). The effect for JpWaC matches the differences that (Sharoff 2006a) found when comparing web corpora with newspaper corpora for English and German. For English, he also considers a third point of comparison: the BNC, a corpus carefully designed to give a balanced picture of contemporary British English and often cited as a model for other corpus-building projects. Sharoff finds that his web corpus (prepared using the same method we

**Table 3** Differences in entries between JpWaC (left table) and News (right table): the most distinctive twenty words, according to the log-likelihood statistic (LL), in each case. Columns JpWaC and News give frequencies per thousand for the two corpora.

word form	tag	LL	JpWaC	News	word form	tag	LL	JpWaC	News
ます	Aux	206122	<b>5.76</b>	0.74	日	N.Suff.ms	110177	1.38	<b>4.34</b>
です	Aux	148844	<b>4.55</b>	0.70	た	Aux	104951	16.95	<b>25.5</b>
まし	Aux	88599	<b>2.69</b>	0.41	同	Pref.N	93209	0.12	<b>1.26</b>
月	N.Suff.ms	65734	<b>1.12</b>	0.00	約	Pref.Num	68451	0.17	<b>1.20</b>
て	P.Conj	64966	<b>21.04</b>	14.54	市	N.Suff.p	64868	0.19	<b>1.20</b>
で	Aux	59324	<b>7.53</b>	3.96	など	P.Adv	64244	1.22	<b>3.25</b>
か	P.advcoordfin	59069	<b>4.90</b>	2.10	容疑	N.g	63021	0.02	<b>0.58</b>
の	N.bnd.g	58270	<b>6.31</b>	3.10	首相	N.g	54740	0.05	<b>0.67</b>
な	Aux	51056	<b>6.63</b>	3.52	円	N.Suff.ms	48315	0.48	<b>1.67</b>
ん	N.bnd.g	42093	<b>1.44</b>	0.27	人	N.Suff.ms	46548	0.80	<b>2.22</b>
こと	N.bnd.g	36013	<b>5.86</b>	3.38	万	N.Num	41995	0.38	<b>1.37</b>
ん	Aux	34973	<b>1.22</b>	0.24	勝	N.Suff.ms	40016	0.01	<b>0.35</b>
ませ	Aux	34454	<b>1.07</b>	0.17	区	N.Suff.p	39628	0.08	<b>0.64</b>
私	N.Pron.g	33163	<b>1.72</b>	0.53	を	P.c.g	38881	21.58	<b>27.2</b>
ござい	Aux	29663	<b>0.55</b>	0.01	東京	N.Prop.p.g	38117	0.24	<b>1.03</b>
その	Adn	28998	<b>2.25</b>	0.93	県	N.Suff.p	37005	0.17	<b>0.85</b>
もの	N.bnd.g	28935	<b>1.80</b>	0.64	で	P.c.g	36297	8.79	<b>12.4</b>
それ	N.Pron.g	28902	<b>1.54</b>	0.48	氏	N.Suff.n	33179	0.28	<b>1.04</b>
よう	N.bnd.Aux	28562	<b>2.76</b>	1.29	後	N.g	33120	0.03	<b>0.40</b>
という	P.c.Phr	28068	<b>2.35</b>	1.02	北朝鮮	N.Prop.p.c	32973	0.05	<b>0.48</b>

have used) is more similar to the BNC, than either of them are to the newspaper corpus. This suggests that JpWaC gives a fuller picture of current Japanese than do newspaper corpora such as Mainichi Shimbun.

The findings are in accordance with several other studies that explore the nature of web corpora. (Kilgarriff and Grefenstette 2003) review the field and show that approved or correct forms are typically orders of magnitude more common on the web than incorrect or disparaged forms. In (Keller and Lapata 2003) they undertake psycholinguistic studies which show that associations between pairs of words as found on the web, using search engine hit-counts, correspond closely to associations in the mental lexicon (and for rarer items, the web associations do a better job of approximating mental-lexicon associations than do extrapolations from the BNC).

The issue of corpus composition is crucial for many potential users of the corpus. Consider for example dictionary publishers, who are perhaps the users with the greatest need for large, balanced corpora. (The impetus for the BNC came from dictionary publishing at Oxford University Press and Longman.) For them, newspaper corpora are inadequate because they include only journalism, and journalism is a relatively formal text type. Dictionaries need to cover the vast number of informal words, phrases and constructions that are common in the spoken lan-

guage but which rarely make it into newspapers. They would like the corpora they use to be not only very large, but also to include a substantial proportion of informal language. In the BNC, this was addressed by gathering and transcribing (at great expense) five million words of conversational spoken English.

Size is also an issue. At 100 million words, the BNC was, in the early 1990s, a vast resource, which lexicographers and other language researchers revelled in. But fifteen years on, it no longer looks large when compared to the web, and lexicographers who have been using it extensively at Oxford University Press and Longman, are vividly aware that for some phenomena, it does not provide enough data. The leading UK dictionary-makers are now all working with larger (and more modern) corpora, of between 200 million and 2 billion words<sup>9</sup>. At 400 million words of heterogenous Japanese, JpWaC is a good size for contemporary dictionary making.

Our initial investigations suggest that JpWaC will be a good corpus, in terms of size, composition, and availability, for Japanese lexicography. For the same reasons it will be a useful resource for exercises such as the preparation of word lists for Japanese language teaching and other research requiring good coverage of current Japanese vocabulary across the full range of genres.

## 2.5 Other Japanese web corpora

We are aware of two Japanese web corpora that are comparable to ours. Ueyama and Baroni (2005) used very similar methods. The differences are simply that their corpus is smaller (at 62 million words) and that they also classify the corpus in terms of topic domains and genre types.

Kawahara and Kurohashi (2006) crawled the Japanese web very extensively and have prepared a corpus around five times larger than JpWaC in form of extracted sentences. The authors kindly allowed us access to the data, and we have examined samples and believe it to be a high quality resource, comparable in many ways to JpWaC.

## 3 The Sketch Engine for Japanese

In this section we describe how JpWaC was prepared for the Sketch Engine and the development of the shallow grammatical relation definitions to support word sketching.

Loading a corpus into the Sketch Engine enables the use of standard functions, such as concordances, which include searching for phrases, collocates, grammatical patterns, sorting concor-

---

<sup>9</sup> See e.g. <http://www.askoxford.com/oec>

dances according to various criteria, identifying “subcorpora” etc. In addition, after defining the grammatical relations and loading them into Sketch Engine, the tool finds all the grammatical relation instances and offers access to the more advanced functions: word sketches, thesaurus, and sketch diffs.

### 3.1 Loading the corpus

The Sketch Engine supports loading of any corpora of any language, either on the command line for local installations of the software, or using the ‘CorpusBuilder’ web interface. In order to create good word sketch results, the corpus should be lemmatized and PoS-tagged. (It is also possible to apply the tool to word forms only, which can still give useful output.)

The input format of the corpus is based on the “word per line” standard developed for the Stuttgart Corpus Tools (Christ 1994). Each word is on a new line, and for each word, there can be a number of fields specifying further information about the word, separated by tabs. The fields of interest here are word form, PoS-tag and lemma. Other constituents such as paragraphs, sentences and documents, can be added with associated attribute-value pairs in XML-like format<sup>10</sup>. An example is given in Figure 2.

As described above, JpWaC had been lemmatized and part-of-speech tagged with the ChaSen tool. It was then converted into word-per-line format.

### 3.2 Preparing grammatical relations

For producing word sketches, thesaurus and sketch diffs grammatical relations need to be defined. This mini-grammar of syntactic patterns enables the system to automatically identify possible relations of words to a given keyword. It is language and tagset-specific. The formalism uses regular expressions over PoS-tags. As an example we give below a simple definition for an adjective modifier relation:

```
=modifies
  2:[tag= ‘‘Ai’’] 1:[tag=‘‘N.*’’]
```

The grammatical relation states that, if we find an adjective (PoS tag Ai) immediately followed by a noun (PoS tag starts with N), then a *modifies* relation holds between the noun (labelled 1) and the adjective (labelled 2).

A more complex definition, as used in JpWaC, is given below. Here we define a pair of relations, *modifies\_Ai* and *modifies\_N*, which are duals of each other: if *w1* is in the relation

<sup>10</sup> Full documentation is available at the Sketch Engine website (<http://www.skechengine.co.uk>).

```

<doc id="http://www.0start-hp.com/voice/index.php">
<s>
月々 月々 N.Adv
2 2 N.Num
6 6 N.Num
3 3 N.Num
円 円 N.Suff.msr
で だ Aux
、 、 Sym.c
あなた あなた N.Pron.g
も も P.bind
プログデビュー プログデビュー Unknown
し する V.free
て て P.Conj
み みる V.bnd
ませ ます Aux
ん ん Aux
か か P.advcoordfin
? ? Sym.g
</s>

```

Fig. 2 An example stretch of the corpus

*modifier\_Ai* to  $w_2$ , then  $w_2$  is in the relation *modifies\_N* to  $w_1$ . We start by declaring that we have two ‘dual’ relations, and give them their names. Then comes the pattern itself, with the two arguments for the two relations labelled 2 and 1. The pattern excludes the *nai* adjectival form (using the operator “!”) and possibly includes a prefix before the noun (using the operator “?”). Since the tag “N.\*” includes also suffixes and bound nouns, we exclude these from the results.

\*DUAL

=modifier\_Ai/modifies\_N

2:[tag=‘Ai.\*’ & word!=‘ない|無い’] [tag=‘Pref.\*’]?

1:[tag=‘N.\*’ & tag!=‘N.Suff.\*’ & tag!=‘N.bnd.\*’]

The Sketch Engine finds all matches for these patterns in the corpus and stores them in a database, complete with the corpus position where the match was located. The database is used for preparing (at run time) word sketches and sketch differences and (at compile time) a thesaurus. When the user calls up a word sketch for a noun, the software counts how often each adjective occurs in the *modifier\_Ai* relation with the noun, and computes the salience of the adjective for the noun using a salience formula based on the Dice co-efficient. If the adjective is

above the salience threshold,<sup>11</sup> it appears in the word sketch.

As mentioned, the Japanese grammatical relations are prepared using the ChaSen PoS tags and tokens. The ChaSen tags are quite detailed (88 tags), and the tokenisation is “narrow”: it splits inflectional morphemes from their stems. This has advantages and disadvantages in the creation of grammatical relations. The advantage is that by using already precisely defined tags and small tokens, it is easier to define desired patterns and the need to specify additional constraints is lower. The main disadvantage is that sometimes a targeted string is divided into several tokens by the analyzer, making it more difficult to define patterns and impossible to retrieve some types of results (for example, 女の人 is divided into 3 tokens, *onna-no-hito*, and is therefore not considered as a unit by the system). However, it is possible to search for this kind of strings in the Concordance window, using the CQL functionality.

We defined 22 grammatical relations for Japanese, mostly using the “dual” type. There is also one symmetric relation (where a match for relation *R* between *w1* and *w2* is also a match for *R* between *w2* and *w1*, and one unary relation (involving only one word). Although the grammatical relations for other languages in the Sketch Engine name relations by their functions, such as subjects, objects, we found it easier to remain on the level of particles (が, は, を etc.) and avoid the complexity of topic vs. subject functions - differences between は and が particles. A similar approach is seen in (Kawahara and Kurohashi 2006).

Since the grammatical relations formalism is sequence-based, it is better suited to languages with fixed word order, such as English. There are already some reports on addressing the problem of free word order in creation of Czech and Slovene word sketches (Kilgarriff et al. 2004; Krek and Kilgarriff 2006), where the simple mechanism of gaps in the patterns is employed as one of the solutions. We also use it for the Japanese word sketches. In the example below, up to 5 tokens are allowed, using the notation  $[ ]\{0,5\}$ , between the case particle で and the corresponding verb.

\*DUAL

=noun で/で verb

2: [tag='N.\*'] [tag='P.c.g' & word='で'] [ ]\{0,5\}

1: [tag='V.\*'] | [tag='N.Vs'] [tag='V.\*']

This kind of pattern covers an example such as:

授業 で , 少し だけ 講演 さ せ て 頂 き ま し た .

N.\* P.c.g& で [ ]\{0,5\} N.Vs V.\*...

<sup>11</sup> The adjective must also occur with the noun above the frequency threshold, and must be one of the N (default = 25) highest-scoring adjectives meeting these criteria.

While for some languages trinary relations (between three dependent items) were useful for identifying prepositional phrases, we defined prepositional phrases employing dual relations. In this way, we were able to specify constraints relevant for a specific particle, which gave higher precision output. On the other hand, to define grammatical relations for the phrasal verbs in Japanese (for example, to query most relevant objects and subjects of idioms, such as S が O を 気に入る) the existing relations proved to be too weak, as they are limited to 3-token relations. While these relations are not displayed in the word sketch, they can be found using the ‘sort collocations’ functionality in the concordancer.

The grammatical relations would ideally be more sophisticated, but we have found that very simple definitions, while linguistically unambitious, produce good results. Linguistically complex instances are missed when using simple definitions, but it is generally the case that a small number of simple patterns cover a high proportion of instances, so the majority of high salience collocates are readily found, given a large enough corpus (Kilgarriff et al. 2004). For Japanese, as for other languages, PoS-tagging errors cause more anomalous output than do weaknesses in the grammar. However these errors are rare and the quality of the word sketch output is good. We evaluated the output of six items (元気, 閉める, 女の子, 多分, 温かい, 温泉) that had all together approx. 1000 collocational instances (grammatical relations’ instances) in the word sketches output. The result showed that only seven output mistakes were due to PoS-tagging errors (for example, 温泉へ行~~う~~ (おこなう) instead of 行~~く~~ (いく)) and four of them were due to a shortcoming of the gramrel specification. Fourteen cases offered valuable collocational information but were incomplete due to ChaSen’s narrow tokenization (っ~~ばい~~の女の子). Nonetheless, in future versions, we plan to add more advanced relations into the system and cover also the instances that are now missed and not able to search for (for example, mutli-word units and ‘suru’ verbs).

### 3.3 Word sketches

The word sketch for a word presents a list of all of its salient collocates, organised by the grammatical relations holding between word and collocate. The grammatical relations are as named and defined in the previous section, and the collocates are as found in the corpus. For each collocate listed, the word sketch provides:

- the statistical salience and frequency with which keyword and collocate occur together;
- links to concordances, so we can explore the pattern by looking at the corpus examples.

The word sketch also provides links to grammatical relations, where we can see how the pattern is defined inside the system.

**閉める** JpWaC freq = 2571

nounを		nounは		coord		bound V		nounに	
	1746 5.0		340 4.0		1370 2.9		660 2.5		169 0.9
ドア	349 9.38	カーテン	7 6.26	閉める	18 6.53	なおす	10 5.13	後ろ手	5 9.16
雨戸	31 8.97	夜間	5 6.11	閉まる	10 6.09	きる	37 4.82	時	18 1.32
カーテン	61 8.83	窓	25 5.51	開ける	22 4.69	おく	39 2.72	ため	17 0.64
扉	135 8.82	扉	11 5.43	閉じる	6 4.0	ちゃう	18 2.64	前	10 0.56
蓋	53 8.41	ドア	13 4.75	忘れる	29 3.55	しまう	105 2.55		
ふた	36 8.3	鍵	7 4.53	消す	5 3.25	もらう	13 1.14		
戸	50 8.29	音段	5 4.32	たる	6 3.24	くれる	19 0.92	<b>suffix</b>	<b>201 0.6</b>
窓	168 8.15	夜	5 2.08	掛ける	6 3.19	てる	33 0.81	っばなし	16 6.44
シャッター	30 8.04	今日	7 1.29	近づく	5 2.73	くださる	8 0.3	させる	21 3.21
元栓	10 7.5	彼	7 1.02	聞こえる	6 2.7	いる	214 0.2	られる	138 2.63
鍵	61 7.46	店	8 0.98	押す	5 2.43	いただく	11 0.12		
蛇口	14 7.46	時	10 0.48	寝る	6 2.02				
バルブ	11 7.23	とき	5 0.37	帰る	11 1.82	<b>nounで</b>	<b>202 2.2</b>		
フタ	13 7.19	場合	6 0.22	かける	14 1.77	後ろ手	6 9.26		
引き戸	9 7.16			える	5 1.75	鍵	6 4.33		
ファスナー	8 7.02	<b>modifier Adv</b>	<b>158 3.9</b>	走る	6 1.56	分	5 0.74		
チャック	7 6.73	きっちり	5 6.34	開く	8 1.51	手	6 0.56		
障子	7 6.63	しっかり	18 5.02	入る	23 1.41				
水門	6 6.57	きちんと	10 4.78	戻る	7 1.21	<b>nounが</b>	<b>153 1.1</b>		
ボンネット	6 6.51	必ず	7 4.74	置く	7 0.98	カーテン	7 6.36		
襖	6 6.51	もう	8 3.95	出来る	11 0.94	扉	10 5.32		
門	24 6.42	ちゃんと	6 3.93	送る	5 0.9	店	7 0.79		
門扉	5 6.4	すぐ	7 3.72	出る	26 0.87				
栓	7 6.36	また	5 3.16	入れる	8 0.71				
サッシ	5 6.2			言う	29 0.37				

Fig. 3 Word sketch for the verb 閉める (*shimeru*, 'to close')

We presented word sketches for the noun お湯 in the opening section; here we give another example, this time for the verb 閉める (*shimeru*) (Figure 3). Here the grammatical relations reveal different noun-particle-verb relations, such as ドアを閉める, カーテンは閉める, カーテン・扉・店が閉められる, 後ろ手・鍵で閉める. We can also easily find the most relevant bound verbs that appear with the verb: 閉めきる, 閉め直す, 閉めてくれる・いただく・もらう etc. Adverbs that usually modify the verb, しっかり, きちんと, 必ず imply that the action is done/should be done firmly, tightly, definitely. The suffixes that appear with the verb (られる, させる, っばなし) suggest the frequent passive and causative usage of the verb. After checking the concordance, we can also select a number of useful usage examples: ドアがきちんと閉められています, 雨戸を閉めさせる, カーテンを閉めっばなしにする).

### 3.4 Thesaurus

The semantic similarity in the Sketch Engine is based on “shared triples” (for example, 雑誌 and 本 share the same triple 〈?を読む〉). When we find a pair of grammatical relation instances, such as 〈雑誌を読む〉 and 〈本を読む〉 with high salience for both words, 本 and 雑誌, we use it as a piece of evidence for assuming the words belong to the same thesaurus category. The thesaurus is built by computing “nearest neighbours” for each word, and based on the tradition of automatic thesaurus building (Sparck 1986; Grefenstette 1994; Lin 1998). We present a thesaurus entry for the word お湯 in Figure 4.

As can be seen from the list of the words in the thesaurus, they can also suggest different senses of a word. In the case of お湯, these are ‘bath’ (風呂) and お湯 as a liquid ‘hot/warm water’ (液体, 熱湯) (see also Section 1). The thesaurus also reveals that the word is semantically most similar to the word 湯 (*yu*), which is actually a variation of the お湯 (*oyu*), and adds on an additional sense ‘hot spring’ (温泉). It also offers sets of related words belonging to the same semantic domains, indicating that お湯 is semantically related to food/drink and its preparation (スープ, 牛乳, お茶, 紅茶, ご飯, 鍋など), to water flow/liquids (液体, 液, 海水, 水道), to bathing (シャワー, 汗, 湯船) etc. It also relates to 水, 熱湯 and its antonym 冷水.

お湯 JpWaC freq = 2876	
湯	0.326 風呂 0.181 温泉 0.091
スープ	0.208 種 0.091 カレー 0.086
牛乳	0.18 ミルク 0.141 紅茶 0.11 飲み物 0.088 ジュース 0.088
海水	0.172
水	0.164 空気 0.116
熱湯	0.163 冷水 0.095
お茶	0.141 コーヒー 0.139 ビール 0.091 ワイン 0.086 酒 0.083
シャワー	0.14 水道 0.078
油	0.122 オイル 0.106 ガソリン 0.085
鍋	0.122 フライパン 0.089
醤油	0.121 塩 0.111 砂糖 0.104
汁	0.121
液体	0.12
洗剤	0.117
湯船	0.103 浴槽 0.086
ご飯	0.102 食べ物 0.096 パン 0.081 野菜 0.08 肉 0.079
火	0.099
血	0.098 汗 0.086
液	0.096

Fig. 4 Thesaurus for the noun お湯 (*oyu*)

### 3.5 Sketch Differences

The difference between two near-synonyms can be identified as the triples that have high salience for one word, but no occurrences (or low salience) for the other. Based on this type of data on various grammatical relations and their salience, a one-page summary for sketch differences between two semantically similar words can be presented. The system is also useful for showing differences in language usage for words that are considered semantically similar but different orthographically, for example 良い (*yoi/ii*) and いい (*ii*), as well as for showing differences of transitive/intransitive semantic pairs, such as 閉める・閉まる (*shimeru/shimaru*).

The sketch difference summary offers the list of collocates that are common for the comparing pair showing their salience and frequency variance, as well as the list of collocates that appear only with one word of the comparing pair.

Figure 5 shows partial results of the sketch difference for 女の子 (*onna no ko*, ‘girl’) and 男の子 (*otoko no ko*, ‘boy’). Ai adjective-noun relations that are common to both (common patterns), and that apply only to one of the words (“女の子” only patterns and “男の子” only patterns) are displayed. From the results we can see that 可愛い, いい, although common to both, is more present as the collocation to 女の子. “Only patterns” reveal that かわいらしい, 美しい, 強い collocate only with 女の子 and that カッコイイ, かっこいい, 弱い appears only with 男の子.

Before we presented word sketches for the word お湯. Comparing it with its coordinate noun 水 in the sketch difference shows clearly that only 水 collocates with 冷たい, おいしい and only お湯 with 熱い.

女の子/男の子 preloaded/jp/wac freq = 16309/6486

Common patterns		"女の子" only patterns		"男の子" only patterns				
女の子	6.0	4.0	2.0	0	2.0	-4.0	-6.0	男の子
modifier_Ai	1021	297	10.0	7.6				
可愛い	178	23	9.4	6.6				
かわいらしい	176	25	9.3	6.6				
若い	288	76	7.9	6.0				
可愛らしい	15	5	7.7	6.7				
幼い	25	17	7.0	6.7				
小さい	41	20	5.5	4.5				
優しい	15	10	5.1	4.6				
つましい	8	7	3.6	3.5				
いい	38	9	2.6	0.5				

  

"女の子" only patterns		"男の子" only patterns	
modifier_Ai	1021	10.0	
思しい	5	7.0	
かわいらしい	7	6.7	
ちっちゃい	5	6.7	
美しい	14	3.8	
明るい	5	3.4	
すごい	8	2.9	
強い	14	2.5	
良い	17	2.0	
悪い	9	1.8	
大きい	9	1.7	
新しい	9	1.6	
高い	8	0.9	

  

"男の子" only patterns	
modifier_Ai	297
カッコイイ	8
かっこいい	10
弱い	5

Fig. 5 Sketch difference for 女の子 and 男の子 (partial)

## 4 Evaluation

We evaluated the word sketches results by comparing them to ten randomly selected entries in the Japanese collocational dictionary 『日本語表現活用辞典』 (Himeno 2004). Here we briefly summarize the results, for a detailed description please refer to (Srdanović and Nishina 2008).

The first difference to be noticed is in the number of grammatical relations holding between the words of a collocation, or collocation types. The Sketch Engine offers a richer set. As well as Ana adjectives and verbs, there are nouns, Ai adjectives, adverbs and others. Also for one word class there is a richer variety of collocations. For example, for verbs, there are not just collocations with が , を , と , に as in the dictionary, but also with で , まで , から , へ and は particles.

The comparison also suggests that the word sketch is a useful tool for selecting the most relevant collocations—for example, the very frequent *かすかな記憶* was missed in the dictionary. There are also some collocations in the dictionary, which were not present (or not regarded as significant) in the JpWaC word sketches. For verbs *閉める* and *閉まる* we find collocations such as *ドア* , *まど* , *カーテン* , but we do not find *襖* and *障子* which are present in the dictionary. This suggests differences in the corpora used (the dictionary uses newspaper corpus, modern literature etc.) and indicates that language change has occurred and these two collocations are no longer current.

In addition, the dictionary usually offers examples for most significant collocations. The Sketch Engine is useful for finding good examples because the user can click on the word in the word sketch to see a concordance of the sentences that the collocation occurs in.

Lastly, Thesaurus and Sketch Diff functions can also be used to easily obtain relevant data on similar words. This type of data can be used for cross-references and to show the differences between similar words, which are rarely captured in the dictionary. Although it seems that *閉める* and *閉まる* share almost the same collocations, some differences can be rapidly found in the Sketch Diff—for example, *レストラン* , *商店* , *図書館* are used only with *閉まる* .

In this evaluation, we concentrated on collocational dictionaries and confirmed that word sketches are a useful tool in compiling such dictionaries. As has already been shown for other languages, the application of the tool is in fact broader—in lexicography, for compilation of various dictionary types, as well as for NLP, language research, and language teaching.

## 5 Conclusion and further work

In this paper we presented how JpWaC, a 400-million-word Japanese web corpus, and a set of Japanese grammatical relations were created and employed inside the Sketch Engine. The Sketch Engine uses grammatical relations (defined with regular expressions over part-of-speech tags) and lexical statistics, applied to a large corpus, to find useful linguistic information: the most salient collocation and grammatical patterns for a word. The tool has already proved to be useful for English and other languages, and we believe that the Japanese version of the tool is a step forward in corpus-based lexicography, language learning, and linguistic research for Japanese. Its possible application in the various fields is investigated and exemplified in (Srdanović and Nishina 2008).

As future work, we plan to make the system user-friendly for both native-speakers and learners of Japanese, by providing a Japanese interface and by offering option to choose between English and Japanese tag sets and grammatical relation names, and by providing the corpus also in furigana and romaji transcriptions. We shall also enrich the grammatical relation set.

We also aim to add some other Japanese corpora into the system, which among other things would be interesting from the point of view of comparing various corpora. Currently a long-term corpus development project is in progress at the National Institute of the Japanese Language (Maekawa 2006). Loading these corpora into the Sketch Engine tool is being considered (Tono 2007). We will also consider possible benefits of implementing some other morphological and structural analysers for the Japanese language. Finally, we shall explore a direct application of the system to the creation of learner's dictionaries (Erjavec et al. 2006) and CALL systems (Nishina and Yoshihashi 2007).

## Acknowledgment

The authors would like to thank Serge Sharoff for providing the URL list, which served as the basis for constructing the JpWaC corpus, all the collaborators of the WAC project for making their software available, and the anonymous reviewers for their useful comments.

## Reference

- Baroni, M. and Bernardini, S. (2004). "BootCat: Bootstrapping corpora and terms from the web." In *Proceedings of the Fourth Language Resources and Evaluation Conference, LREC2004* Lisbon.

- Baroni, M. and Bernardini, S. (Eds.) (2006). *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna.
- Baroni, M. and Kilgarriff, A. (2006). "Large linguistically-processed Web corpora for multiple languages." In *Proceedings EACL* Trento, Italy.
- Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). "WebBootCaT: instant domain-specific corpora to support human translators." In *Proceedings of EAMT 2006*, pp. 47–252 Oslo.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.
- Biber, D. (1995). *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.
- Chantree, F., de Roeck, A., Kilgarriff, A., and Willis, A. (2005). "Disambiguating Coordinations Using Word Distribution Information." In *Proceedings RANLP* Bulgaria.
- Chen, A., Rychlý, P., Huang, C.-R., Kilgarriff, A., and Smith, S. (2007). "A corpus query tool for SLA: learning Mandarin with the help of Sketch Engine." In *Practical Applications of Language Corpora (PALC)* Lodz, Poland.
- Christ, O. (1994). "A modular and flexible architecture for an integrated corpus query system." In *COMPLEX' 94* Budapest.
- Erjavec, T., Hmeljak, K. S., and Srđanović, I. E. (2006). "jaSlo, A Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement." In *Proceedings of the 12th EURALEX International Congress* Turin, Italy.
- Gatt, A. and van Deemter, K. (2006). "Conceptual coherence in the generation of referring expressions." In *Proceedings of the COLING-ACL 2006 Main Conference Poster Session*.
- Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer.
- Heylighen, F. (2002). "Variation in the Contextuality of Language: An Empirical Measure." *Foundations of Science*, **7** (3), pp. 293–340.
- Himeno, M. (2004). *Nihongo hyougen katsuyou jiten*. Kenkyusha.
- Kawahara, D. and Kurohashi, S. (2006). "Case Frame Compilation from the Web using High-Performance Computing." In *Proceedings LREC* Genoa, Italy.
- Keller, F. and Lapata, M. (2003). "Using the Web to Obtain Frequencies for Unseen Bigrams." *Computational Linguistics*, **29** (3), pp. 459–484.
- Kilgarriff, A. (2001). "Comparing Corpora." *International Journal of Corpus Linguistics*, **6** (1), pp. 1–37.
- Kilgarriff, A. and Grefenstette, G. (2003). "Introduction to the Special Issue on Web as Corpus." *Computational Linguistics*, **29** (3).

- Kilgarriff, A. and Rundell, M. (2002). “Lexical profiling software and its lexicographic applications—a case study.” In *Proceedings EURALEX*, pp. 807–818 Copenhagen.
- Kilgarriff, A., Rychly, P., Smrž, P., and Tugwell, D. (2004). “The Sketch Engine.” In *Proceedings EURALEX*, pp. 105–116 Lorient, France.
- Krek, S. and Kilgarriff, A. (2006). “Slovene Word Sketches.” In *Proceedings 5th Slovenian/First International Languages Technology Conference* Ljubljana, Slovenia.
- Lin, D. (1998). “Automatic retrieval; and clustering of similar words.” In *Proceedings COLING-ACL*, pp. 768–774 Montreal.
- Maekawa, K. (2006). “Kotonoha. The Corpus Development Project of the National Institute for Japanese Language.” In *Proceedings of the 13th NIJL International Symposium: Language Corpora: Their Compilation and Application*, pp. 55–62 Tokyo.
- Nishina, K. and Yoshihashi, K. (2007). “Japanese Composition Support System Displaying Occurrences and Example Sentences.” In *Symposium on Large-scale Knowledge Resources (LKR2007)*, pp. 119–122.
- Rayson, P. and Garside, R. (2000). “Comparing corpora using frequency profiling.” In *Proceedings of the ACL Workshop on Comparing Corpora*, pp. 1–6 Hong Kong.
- Sharoff, S. (2006a). “Creating general-purpose corpora using automated search engine queries.” In *WaCky! Working papers on the Web as Corpus*. GEDIT, Bologna.
- Sharoff, S. (2006b). “Open-source corpora: using the net to fish for linguistic data.” *International Journal of Corpus Linguistics*, **11** (4), pp. 435–462.
- Sparck, K. J. (1986). *Synonymy and Semantic Classification*. Edinburgh University Press.
- Srdanović, I. E. and Nishina, K. (2008). “Ko-pasu kensaku tsu-ru Sketch Engine no nihongoban to sono riyō hōhō (The Sketch Engine corpus query tool for Japanese and its possible applications).” *Nihongo kagaku (Japanese Linguistics)*, **24**, pp. 59–80.
- Tono, Y. (2007). “Nihongo ko-pasu de no Sketch Engine jissō no kokoromi (Using the Sketch Engine for Japanese Corpora).” In *Tokutei riyōiki kenkyū “Nihongo ko-pasu” Heisei 18 nendo kōkai wa-kushoppu (Kenkyū seika hōkokukai) yokōshū*. Monbukagakusho kagaku kenkyūhi tokutei riyōiki kenkyū “Nihongo ko-pasu”, pp. 109–112 Soukatsuhan.
- Ueyama, M. and Baroni, M. (2005). “Automated construction and evaluation of a Japanese web-based reference corpus.” In *Proceedings of Corpus Linguistics 2005* Birmingham.
- Weeds, J. and Weir, D. (2005). “Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity.”

**Irena Srdanović Erjavec:** received the Bachelor degree in Japanese Language from University of Belgrade in 1997, and Master degree in Linguistics from University of Ljubljana in 2005. Since April 2007 she has been a PhD student at Human System Science Department at Tokyo Institute of Technology. From 1997 to 2001, she worked as Japanese language advisor and technical writer at Hermes SoftLab in Ljubljana, from 2001 to 2002, as a translator and international trainee coordinator at Pioneer Corporation in Tokyo, and from 2005 to 2006 as a teacher assistant for Japanese Language at University of Ljubljana. Her research interests lie in the fields of corpus linguistics, lexicography and Japanese language education, particularly concentrating on application of human language technologies.

**Tomaz Erjavec:** received his BSc (1984), MSc (1990), and PhD (1997) degrees in Computer Science from the University of Ljubljana; he also received an M.Sc. in Cognitive Science (1992) from the University of Edinburgh. He works as a scientific associate at the Dept. of Knowledge Technologies at the research Institute Jožef Stefan, with teaching positions at several universities. He has been a visiting researcher at the University of Edinburgh, University of Tokyo and the Joint Research Center of the European Commission. His research interests lie in the fields of computational linguistics and language technologies, with a large part of the work devoted to developing Slovene and multilingual language resources. He has served two terms on the Board of the EACL, has been a member of the Text Encoding Initiative Council and was the founding president of the Slovenian Language Technologies Society. See also <http://nl.ijs.si/et/>

**Adam Kilgarriff:** is Director of Lexical Computing Ltd. which has developed the Sketch Engine <http://www.sketchengine.co.uk>, a leading tool for corpus research. His scientific interests lie at the intersection of computational linguistics, corpus linguistics, and dictionary-making. Following a PhD on “Polysemy” from Sussex University, he has worked at Longman Dictionaries, Oxford University Press, and the University of Brighton, and is now Director of the Lexicography MasterClass (<http://www.lexmasterclass.com>) as well as Lexical Computing Ltd. He is Visiting Research Fellow at the Universities of Leeds and Sussex. He started the SENSEVAL initiative on automatic word sense disambiguation and is now active in moves to make the web available as a linguists’

corpus. He is the founding chair of ACL-SIGWAC (Association for Computational Linguistics Special Interest Group on Web as Corpus) and has been chair of ACL-SIG on the lexicon and Board member of EURALEX (European Association for Lexicography). See also <http://www.kilgarriff.co.uk/>

(Received July 3, 2007)

(Revised October 23, 2007)

(Accepted November 15, 2007)