



Japanese word sketches: towards a new version

Irena Srdanović

irena.srdanovic@gmail.com

Overview

- Japanese word sketches (intro)
 - Jap gramrel & ChaSen tagset specifics
- Evaluations:
 - Comparing to Jap collocational dictionary
 - SketchEval project
- Next version
- Sub-corpus & distant collocations
- Web corpus vs. balanced corpus

Japanese corpus linguistics



■ Before:

- Aozora bunko (literal texts)
- newspaper data (commercial use)
- various corpora used inside an institution...

■ From 2005:

- 5-year project at National Institute for Japanese Language (Balanced Corpus of Japanese...)
- > 2007, Web corpus into SkE (400 million tokens)

Steps for JpWaC (Erjavec et al 2007)

- URL list of pages in Japanese
 - provided by S. Sharoff
- Files downloaded and cleaned with BootCat
 - BootCat created by M. Baroni and others from the WaCky project, c.f. <http://wacky.sslmit.unibo.it/>
- Segmented, tokenised, tagged with ChaSen
 - By T. Erjavec, ChaSen available at <http://chasen.naist.jp/hiki/ChaSen/>
- Translated ChaSen tags to English
 - by Srdanovic, also used in the jaSlo dictionary project (Hmeljak Sangawa et al)
- Converted to Sketch Engine format and loaded

ChaSen morphological analyzer

- 88 tags
- classification of some POS categories is very detailed
- suffixes, prefixes included

<u>suffix</u>	<u>52733</u>	<u>5.4</u>
室	<u>8915</u>	11.26
科	<u>5545</u>	11.09
員	<u>4232</u>	9.95
会	<u>11873</u>	9.78
費	<u>1834</u>	8.92
者	<u>14911</u>	8.85

- -shitsu (research lab)
- -ka (research department)
- -in (research member)
- -kai (society)
- -hi (research expenses)
- -sha (researcher)

Word sketch example

研究 JpWaC freq = 178341

suffix 52733 5.4	modifier Ana 2340 3.0	pronomの 27232 2.9	prefix 3268 1.9	をverb 18277 1.9
室 8915 11.26	優秀 105 8.27	大学 682 7.88	当 270 9.08	進める 1286 8.46
科 5545 11.09	重点的 18 7.72	分野 521 7.84	本 1891 8.55	重ねる 179 7.21
員 4232 9.95	地道 23 7.62	最近 378 7.43	各 275 7.19	行う 1812 7.0
会 11873 9.78	アカデミック 18 7.56	最先端 149 7.33	同 147 6.65	すすめる 70 6.49
費 1834 8.92	著名 30 7.44	今後 311 7.17	準 21 6.52	続ける 464 6.42
者 14911 8.85	更 27 7.37	研究所 161 7.12	諸 52 6.15	行なう 162 6.38
院 321 7.17	広範 21 7.13	最新 179 7.11	某 29 6.08	おこなう 68 6.14
棟 202 6.87	主要 59 7.0	多く 436 6.92	御 138 5.69	始める 340 6.12
官 282 6.71	活発 35 6.96	技術 408 6.81	新 65 4.94	深める 61 5.94
家 903 6.67	ユニーク 24 6.94	史 188 6.8	大 110 4.66	まとめる 75 5.24
職 226 6.64	画期的 19 6.9	学 378 6.74	元 34 4.31	志す 23 5.24
部 408 6.4	さまざま 107 6.83	樹 101 6.67	全 18 3.79	踏まえる 42 5.14

modifier Ai 1505 1.8	のpronom 11222 1.2	particle 6353 1.2	がverb 3764 1.0	がAdj 963 0.9
幅広い 48 7.37	成果 588 8.94	を通して 100 7.24	進む 454 7.32	盛ん 55 9.06
若い 176 7.21	第一人者 149 8.64	を通じて 97 7.02	盛る 30 6.72	さかん 15 8.46
興味深い 40 6.85	進展 141 7.95	において 431 6.93	なす 158 6.27	活発 36 7.24
詳しい 47 6.54	推進 178 7.75	に関する 311 6.85	すすむ 14 6.09	数多い 14 6.57

Gramrel example

```
# JAPANESE word sketches ver. 0.3
# Made by Irena
# 2007-03-13

*STRUCTLIMIT s
*DEFAULTATTR tag

*SYMMETRIC
=coord
  1:"N.*" "P.coord" "Pref.*"? 2:[tag="N.*" & tag!="N.Num"]
  1:"V.*" []{0,5} 2:"V.free"
  1:"Ai.*" []{0,3} 2:[tag="Ai.free" & word!="ない|無い"] | [tag="N.Ana"]
  1:"N.Ana" []{0,3} 2:[tag="Ai.free" & word!="ない|無い"] | [tag="N.Ana"]

*DUAL
=modifier_Ai/modifies_N
  2:[tag="Ai.*" & word!="ない|無い"] [tag="Pref.*"?
  1:[tag="N.*" & tag!="N.Suff.*" & tag!="N.bnd.*"]

*DUAL
=modifier_Ana/modifies_N
  2:"N.Ana" "Aux" "Pref.*"? 1:[tag="N.*" & tag!="N.Suff.*" & tag!="N.bnd.*"]

*DUAL
=Nが/がAdj
  2:[tag="N.*"] [tag="P.c.g" & word="が"] []{0,2} 1:"Ai.*|N.Ana"

*DUAL
=Nは/はAdj
  2:[tag="N.*"] [tag="P.bind" & word="は"] []{0,2} 1:"Ai.*|N.Ana"
```

- (Srdanovic et al 2008)
- 22 relations, mainly “dual”, one “symmetric”, one “unary”
- Names not always by functions
- formalism is sequence based -> mechanism of gaps []{0,5}

Covered collocational relations (1)

Nouns

PoS	Gramrel relation	Type of relation	Example
Noun	modifier_Ai	Adj Ai modifying noun	新しい挑戦
	modifier_Ana	Adj Ana modifying noun	果敢な挑戦
	をverb	Nwo + verb	挑戦を受ける
	でverb	Nde + verb	お湯で溶く
	がverb	Nga + verb	挑戦が始まる
	にverb	Nni + verb	挑戦に立ち向かう
	はverb	Nwa + verb	挑戦は続く
	からverb	Nkara + verb	お湯から上がる
	pronomの	noun + noN	最後の挑戦
	のpronom	Nno + noun	挑戦の意欲
	がAdj	Nwa + Adj	お湯がいい
	はAdj	Nga + Adj	お湯はぬるい
	coord	coordinate relation	挑戦・革新
	particle	N + particle	挑戦という
	suffix	N + suffix	挑戦状
	prefix	prefix + N	初挑戦

Covered collocational relations (2)

Verbs

PoS	Gramrel relation	Type of relation	Example
Verb	modifier_Adv	Adv modifying V	ここにこ笑う
	nounは	noun_wa + V	彼は笑う
	nounが	noun_ga + V	鬼が笑う
	bound_V	bound verbs connecting to free verbs	わらっちゃう
	V_bound	free verbs connected to bound verbs	連れて行く
	nounで	noun_de + V	鼻で笑う
	nounに	noun_ni + V	高らかに笑う
	nounから	noun_kara + V	(心の)低から笑う
	nounまで	noun_made + V	最後まで笑う
	nounを	noun_wo + V	腹を(抱えて)笑う
	nounへ	noun_he + V	(公園へ行く)
	coord	coordinate relation	笑う・泣く
	suffix	V+suffix	笑いっぱなし
	prefix	prefix + V	超笑う

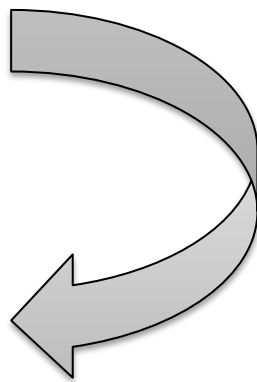
Covered collocational relations (3)

Adjectives Ai/Ana, Adverbs

PoS		Gramrel relation	Type of relation	Example
Adj Ai	7	modifies_N	Ai modifies noun	長い歴史
		Nは	Nwa + Ai	道のりは長い
		Nが	Nga + Ai	前置きが長い
		bound_N	bound/free nouns connecting to Ai	長いわけ
		coord	coordinate relation	長い・短い
		suffix	Ai+suffix	超長い
		prefix	prefix + Ai	長さ
Adj Ana	11	bound_N	bound/free nouns connecting to N.Ana	重要な点
		Nは	Nwa + N.Ana	役割は重要
		Nが	Nga + N.Ana	ことが重要
		pronomの	noun + no N.Ana	重要なネットワーク
		の pronom	N.Ana_no + noun	重要な課題
		modifier_Ai	Ai modifying N.Ana	ものすごい重要
		modifier_Ana	N.Ana modifying N.Ana	不可欠な重要(な)
		suffix	N.Ana+suffix	重要性
		prefix	prefix + N.Ana	最重要
		coord	coordinate relation	重要・高い
		particle	N.Ana + particle	重要と(なる・いう)
Adv	1	modifies_V	Adv modifying V	やっと落ち着く

Number of types & tokens covered

PoS	Token	Type
noun	139,864,402	121,732
verb	39,630,531	17,641
Adj	4,119,406	1,863
Adv	6,141,530	2,820



PoS	ChaSen Tags (eng)	Token	Type
Noun	N.g	50,121,542	49,268
	N.Vs	21,704,856	10,017
	N.Ana	4,436,970	2,892
	other	63,601,034	59,555
	total	139,864,402	121,732

Evaluation

◎ Evaluation 1 : Comparing with collocational dictionary for language learners

- “*Nihongo hyougen katsuyou jiten*” (Himeno 2004)
- 10 entries for verbs and adjectives –na (Ana)

◎ Evaluation 2 : SketchEval

- *is this word a good candidate for inclusion in the headword's collocation-dictionary entry?*
- nouns, adjectives, verbs (2:1:1 ratio)
(42items×20 collocations)

- Suru verbs as nouns N.Vs
- Ana adjectives as nouns N.Ana

Results of the Evaluation 1 (part)

- ① We can extract much more types of collocational relations by SkE than the dictionary covers;
 - we can decide on the most salient collocations
- Dictionary: covers only collocations of verbs and adjectives –na (Ana)
- Dictionary (verbs): Noun + *ga, wo, to, ni* + verb
- SkE(verbs) : Noun + *ga, wo, to, ni, de, made, kara, he...* + verb, coordinate relations with other verbs, collocating with adverbs, bound verbs etc.
- ② Most salient & frequent collocations in Jap word sketches not necessarily present in the dictionary (*kasukana kioku* etc.)

Results of the Evaluation 2

Selection choice	Answer			
	EvalA	EvalB	EvalC	Avarage
Good	83.40%	81.98%	38.93%	68.10%
Good (but wrong grammatical relation)	2.27%	0.46%	2.50%	1.74%
Maybe (not so striking collocate)	9.69%	3.72%	40.23%	17.88%
Maybe (specialized vocabulary)	0.31%	1.51%	0.60%	0.81%
Bad	3.30%	12.33%	17.74%	11.12%
N/A	1.03%	0%	0%	0.34%

avarage for
high freq.
words :
Good 76.37%

	Evaluated items (747)	
	Three agree (294)	Two agree (690)
Good	278 (94.5%)	600 (86.95%)
Bad	16 (5.5%)	90 (13.05%)

Problem of “incomplete collocations”

- “Good but not complete”
- Comes from detailed ChaSen tagset
- researcher: *kenkyu* + *sha*

research + er

extensive research ≠ extensive researcher

- girl: *onna* + *no* + *ko*
woman + poss + child

little girl ≠ little woman

To solve the problem
try UniDic/McCab!

Some misses in the current WS

- “suru” verbs don’t appear as collocates
 - where other types of verbs appear, since they are tagged as nouns (“N.Vs”)
(for example, Adv + Verb doesn’t cover “suru” verbs)
- Compound nouns are not covered in the current gramrel (N+N)

To add in the
next version!

Corpus & salience

- Corpus related problems

- Duplicates: when the same pages (or their copies) appear a number of times

Corpus clean-up!

- Salience related problems

- When some collocate appears very frequently but only from one source (one web page)

To find a way to exclude
this kind of cases! Espec. relevant
for web corpora!

(Distant) collocations

- “You shall know a word by the company it keeps” (First)
- ”collocation is the occurrence of two or more words within a short space of each other in a text”
(usually referred to 5 words at most) (Sinclair)
- “words that co-occur more often than chance”
 - MI: extracting pairs of correlated words (collocations) within a fixed distance of 5 words
- Notion of “distant collocation” only recently
 - For extracting collocations interrupted by a string or two, usually within a short distance
 - “interrupted collocations”, “discontinuous collocations”

Kitto Tanaka-san no otousan wa ashita ka asatte kuru hazu da.
Adverb-----Modality form

Extracting Adverbs and Clause-Final Modality Distant Collocations

Adverbs + distant collocations

- verbs
- adjectives
- final particles

Recognized by ChaSen
→
simply add new relations into the gramrel file

Adverbs + (distant) modality forms

1. Create comprehensive list of modality forms and variations
2. Define ChaSen units form modality forms and create a new “Mod” tag
3. Retag the corpus (add “Mod” tag)
4. Add a new relation into the Gramrel file (Srdanovic et al 2009)

not available in ChaSen

→

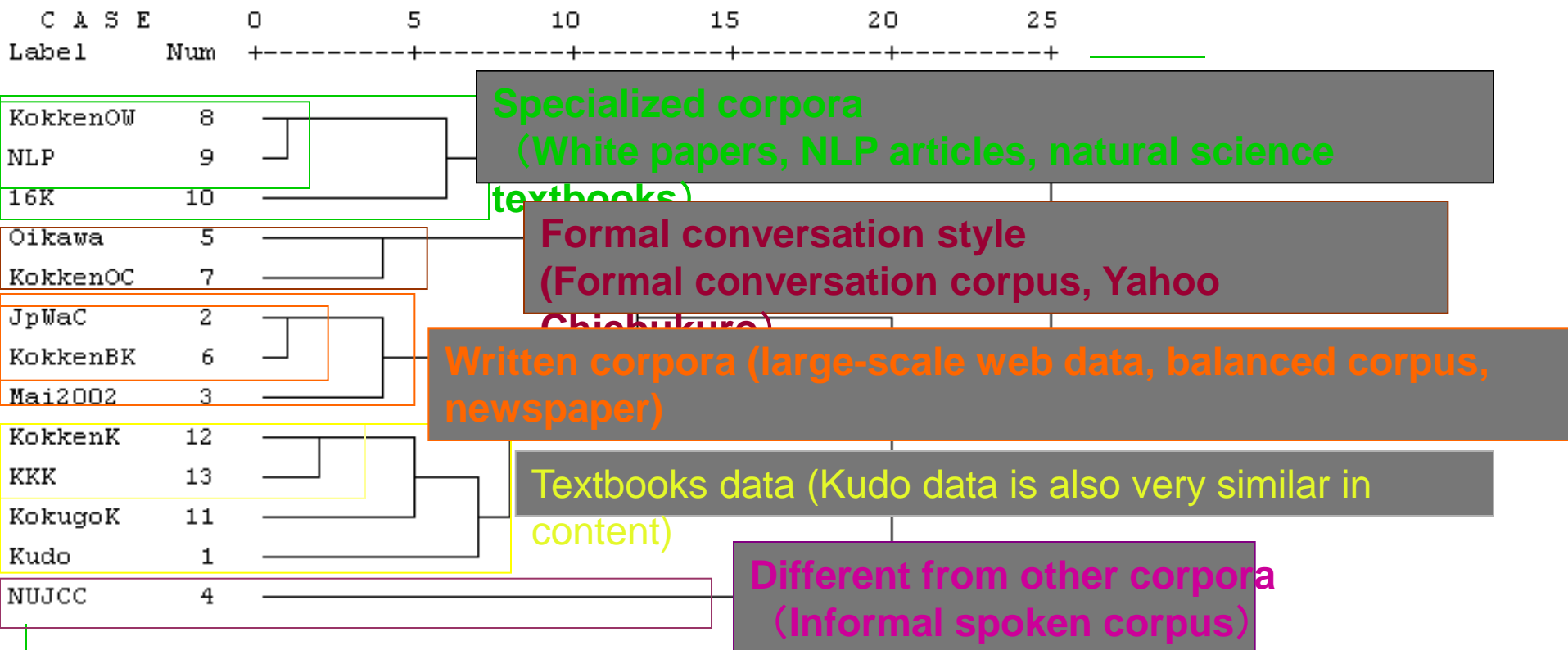
next 4 steps

Modality forms and variations

Before retag			→	After retag		
Token	Lemma	POS		Token	Lemma	POS
<i>kamo</i>	<i>kamo</i>	P. Adv		<i>kamoshirenai</i>	<i>kamoshirenai</i>	Mod
<i>shire</i>	<i>shiru</i>	V. free				
<i>nai</i>	<i>nai</i>	Aux				

- Variations (inflection, style, orthography: kanji or kana)
 - kamoshiremasen, kamoshirenai, kamoshiren, kamoshirenu
- Combined modality forms
 - toomou kamoshirenai, toomou no kamoshirenai, kamoshirenai noda
- Number of modality forms
 - Basic modality forms 31
 - Combined modality forms 596
 - Variations 2641
- Evaluation: very good results! 93 – 96% of accuracy

Corpus classification based on adverb distribution



Extracted collocations of adverbs & modality forms (web corpus)

MOD	Adverb	Freq	Most frequent clause-final modality forms (frequency & salience)				
EXP	tabun	1527	のだろう 124 3659、と思う 221 3587、だろう 175 3523、のだと思う 32 3022、のではないかと 28 2482				
NEC/EXP	kanarazu	1448	はず 31 2809、のだ 95 2768、と思う 41 2196、だろう 20 1629、わけだ 6 1157				
EXP	osoraku	1341	だろう 269 4121、のだろう 128 367、ことだろう 37 2906、と思う 131 2838、に違いない 21 260				
EXP/NEC	kitto	1340	のだろう 160 4104、だろう 159 3454、に違いない 34 3309、ことだろう 44 3213、はず 52 2905				
NEC/EXP	zettai	1294	のだ 87 2535、だろう 40 2256、と思う 41 2092、はず 17 2012、べき 13 1666				
CON	doumo	1022	ようだ 59 344、らしい 44 3365、気がする 28 2732、のだ 63 2216、のようだ 12 2152				
NEC	zattaini	816	てはならない 21 3494、のだ 79 2643、べき 15 1949、と思う 29 1908、はず 12 1788				
CON	douyara	548	らしい 121 4854、ようだ 95 4061、のようだ 27 3121、みたいだ 16 262、そうだ 21 2146				
NEC/EXP	taitei	497	のだ 45 2371、と思う 17 1679、だろう 14 1636、はず 6 1351、ように思う 3 999				
CON	yohodo	409	のだろう 34 2812、だろう 22 1887、と思う 25 1855、のだ 31 1726、のか 17 1645				
NEC	kanarazushimo	382	とは限らない 53 5105、わけではない 35 3477、ものではない 11 2189、ことではない 9 2109、といえない 8 2093				
POSS	moshikashitara	316	かもしれない 102 4336、のかもしれない 59 3967、のかな 10 1906、のではないかと 8 1587、のではないかと 5 1204				
POSS	angai	187	のかもしれない 11 2214、気がする 9 1942、のだ 18 1613、かもしれない 8 161、だろうかと思う 2 1553				
POSS	hyottoshitara	81	かもしれない 25 3048、のかもしれない 17 2867、と考えるのかもしれない 1 1025、のではないかと 2 816、かもしれないのだ 1 787				
NEC/EXP	taigai	72	のだ 6 1116、と思う 4 993、わけだ 2 88、のかなと思う 1 871、だろう 3 869				
EXP	sazo	31	ことだろう 6 205、だろう 8 1639、のだろう 5 1441、だろうと思う 2 1159、だろうと考える 1 916				
EXP	ookata	24	ものと思う 1 846、ものだろう 1 777、と思う 2 742、のかな 1 672、と考える 1 652				
(CON)	kotoniyoruto	2	らしい 1 746、べき 1 702				

● EXP & NEC are most frequent

→ EXP & NEC have functionally greater priority than CON & POSS in Japanese language communication (Srdanovic et al 2009)

Extracted collocations of adverbs & modality forms (balanced corpus)

- Similar results as in web data
- EXP & NEC are most frequent

MOD	Adverb	Freq	Most frequent clause-final modality forms (freq)									
NEC/EXP	kanarazu	4548	のだ	334	はずだ	163	だろう	151	と思う	76	に違いない	44
EXP	osoraku	4216	だろう	961	のだろう	625	と思う	225	に違いない	157	のだ	153
EXP/NEC	kitto	3547	だろう	417	のだろう	332	に違いない	251	と思う	224	はず	157
EXP	tabun	3241	だろう	487	のだろう	412	と思う	290	のだ	111	のではない	109
CON	doumo	2320	のだ	156	らしい	149	ようだ	126	気がする	65	と思う	43
NEC	zettaini	2114	のだ	141	はずだ	61	だろう	59	と思う	51	べきだ	25
POSS	moshikashitara	1824	かもしれない	401	のかもしれない	314	のではない	130	のか	95	のだ	70
NEC	kamoshirenai	1591	わけではない	127	のだ	97	ものではない	69	とはいえない	78	とは限らない	72
NEC/EXP	zettai	1581	のだ	111	と思う	58	だろう	38	はずだ	26	のか	25
CON	douyara	1512	らしい	504	ようだ	319	のようだ	87	のだ	42	だろう	12
NEC/EXP	taitei	1217	のだ	98	だろう	29	ことではない	18	ようだ	15	と思う	15
EXP	yohodo	1047	のだろう	95	のか	76	のだ	60	だろう	43	らしい	42
POSS	hyottoshitara	967	かもしれない	218	のかもしれない	148	のではない	111	のか	36	と思う	36
EXP	sazo	427	だろう	128	ことだろう	76	のだろう	29	に違いない	26	と思う	23
POSS	angai	423	のだ	53	のかもしれない	35	かもしれない	35	のではない	8	ではない	7
EXP	ookata	281	のだろう	31	のだ	21	だろう	16	と思う	10	ことだろう	6
NEC/EXP	taigai	172	のだ	11	だろう	2	だろう	2	はずだ	2	のか	2
POSS	kotoniyoruto	62	のかもしれない	10	かもしれない	8	のではない	4	のか	3	ものであろう	2

Conclusion

- Jap word sketches specifics
 - ChaSen tagset is very narrow -> very detailed results but “incomplete collocations” problem
 - 22 gramrel -> ~50 types of relations
- Evaluation results very good, but as future tasks:
 - “suru” verbs, compound nouns, corpus clean-up, double tagset, proficiency levels
- Adverb-Modality distant collocations
 - sub-corpus, retag, new gramrels
 - in future more of this kind of info
- Web corpus gives balanced results

References

- Srdanović, I., Hodošček B., Bekeš, A., Nishina, K. (2009) "Uebu ko-pasu to kensaku shisutemu wo riyō shita suiryō fukushi to modariti keishiki no enkaku kyouki chuushutsu to nihongo kyouiku he no ouyou", *Shizen gengo shori* (Extracting distant collocations of adverbs and modality forms using web corpus and query system , *Journal of Natural Language Processing*), 16/4, 29-46
- Srdanović, I., Bekeš, A., Nishina, K. (2009) "Ko-pasu ni motozuita goi shirabasu sakusei ni mukete: suiryōteki fukushi to bunmatsu modariti no kyouki wo chuushin ni shite", *Nihongo kyouiku*, (Towards corpus-based creation of lexical syllabus: collocations between suppositional adverbs and clause-final modality forms, *Journal of Japanese Language Education*), 142, 69-79
- Srdanović, E.I., Erjavec, T., Kilgarriff, A. (2008) "A web corpus and word-sketches for Japanese", *Shizen gengo shori (Journal of Natural Language Processing)* 15/2, 137-159
- Srdanović, E.I., Erjavec, T., Kilgarriff, A. (2008) "A web corpus and word-sketches for Japanese", *Information and Media Technologies* 3/3, 2008, 529-551, reprinted from *Journal of Natural Language Processing* 15/2, 137-159
- Srdanović, E.I., Nishina, K. (2008) "Ko-pasu kensaku tsu-ru Sketch Engine no nihongoban to sono riyō houhou", *Nihongo kagaku* (The Sketch Engine corpus query tool for Japanese and its possible applications, *Japanese Linguistics*) 23, 59-80
- Erjavec, T., Srdanović, I., Kilgarriff, A. (2007) A large public-access Japanese corpus and its query tool, CoJaS 2007, *The Inaugural Workshop on Computational Japanese Studies*, March 15-16 2007, Ikaho
- Sharoff, S. (2006) "Creating general-purpose corpora using automated search engine queries." In WaCky! Working papers on the Web as Corpus. GEDIT, Bologna.
- Sharoff, S. (2006) "Open-source corpora: using the net to fish for linguistic data." *International Journal of Corpus Linguistics*, 11 (4), pp. 435-462.
- Erjavec, T., Hmeljak, K. S., and Srdanovic, I. E. (2006) "jaSlo, A Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement." In *Proceedings of the 12th EURALEX International Congress* Turin, Italy.
- 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.

Corpora used: thirteen Japanese corpora of various types

Corpus type		Abbrev.	Explanation	Accessibility
Spoken		NUJCC	Informal conversations (3,584KB)	Nagoya University http://tell.fl.purdue.edu/chakoshi/
		Oikawa	Formal interviews (817KB)	Sokendai
Web		KokkenOC	Yahoo! Chiebukuro, bulletin-board consisting of questions and answers (BCCWJ's sample 2007) (16.3MB)	National Institute for Japanese Language
		JpWaC	Large-scale web corpus (7.3GB)	Japanese version of Sketch Engine http://www.sketchengine.com
Written	Textbooks	KokugoK	Japanese language textbooks for elementary school (3, 723KB)	Tokyo Institute of Technology
		KokkenK	Textbooks for secondary school (BCCWJ) (4.37MB)	National Institute for Japanese Language
		KKK	Japanese language textbooks contained in KokkenK (788KB)	National Institute for Japanese Language
	Natural science textbooks	16K	16 textbooks for natural science university students (2.45MB)	Tokyo Institute of Technology
	Natural science papers	NLP	Japanese NLP journal papers (719KB)	Japanese Society for NLP
	White papers	KokkenOW	Governmental white papers (BCCWJ's sample) (16.4MB)	National Institute for Japanese Language
	Balanced corpus	KokkenBK	Books, periodicals, magazines, newspapers from "publication" and "library" (BCCWJ's sample) (68.6MB; 140MB)	National Institute for Japanese Language
	Newspaper	Mai2002	Mainichi shinbun, newspaper data for the year 2002 (100MB)	Mainichi Shinbun (CD-ROM)
	Other	Kudo	Newspapers, modern literature	Kudo (2000)

Distribution of adverbs in corpora

Adverb/Corpus	KokkenOW	NLP	16K	NUJCC	Oikawa	KokkenOC	JpWaC	KokkenBK	Mai2002	KokugoK	KokkenK	KKK	Kudo
kanarazu	5%	23%	42%	7%	14%	4%	8%	15%	25%	12%	28%	16%	4%
zettai	2%			52%		14%	9%	6%	11%	3%	2%	4%	5%
zettaini	2%		4%			11%	6%	8%	12%	3%	9%	2%	
kanarazushimo	84%	66%	39%	1%	5%	2%	6%	6%	8%	0%	10%	6%	
yohodo	0%				1%	2%	2%	3%	2%	3%	1%	2%	4%
yoppodo				2%		2%	1%	1%	1%	1%			
taigai				2%	8%	1%	1%	1%	0%				1%
taitei	1%	6%	4%	1%		5%	4%	2%	3%	4%	6%	12%	1%
kitto		3%		15%	8%	15%	12%	14%	10%	38%	26%	26%	28%
ookata					1%	0%	0%	1%	0%	1%	1%	2%	3%
osoraku	1%	3%	7%	1%	8%	1%	13%	12%	9%	2%	5%	10%	19%
sazo						0%	0%	1%	1%	4%	1%	2%	5%
tabun	2%		3%	3%	39%	26%	16%	11%	6%	3%	4%	8%	10%
doumo	0%		1%	6%	7%	6%	8%	7%	5%	15%	2%	4%	5%
douyara				2%		3%	5%	5%	3%	3%			5%
angai	0%		1%	3%	1%	0%	2%	1%	1%	1%	1%		2%
hyottoshitara				1%	1%	1%	1%	1%	1%	1%	1%	2%	3%
hyottosuruto							0%						
kotoniyoreba						0%	0%	0%	0%				
kotoniyoruto	3%		1%				0%	0%	1%				1%
moshikashitara				5%	8%	5%	3%	3%	1%	2%	2%	2%	5%
moshikasureba						1%	0%						
moshikasuruto							1%	1%	0%	1%	1%	2%	
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Imbalanced distribution :

- KokkenOW (white papers),
- NLP articles,
- 16K (natural science textbooks),
- NUJCC (informal conversation)

Balanced distribution :

- JpWaC (large-scale web corpus),
- KokkenBK (books)