A large public-access Japanese corpus and its query tool

- JapWaC and Sketch Engine -

Tomaž Erjavec[1], Adam Kilgarriff[2], Irena Srdanović Erjavec[3]

[1]Jožef Stefan Institute, Slovenia
[2]Lexical Computing Ltd. and University of Leeds, UK
[3]Tokyo Institute of Technology, Japan

Overview

1. The case for corpora
2. The case for web corpora
3. How JapWaC was created
4. Sketch Engine (SkE)
5. Demo-ing JapWaC & SkE
6. Future work
7. Access to JapWaC & SkE

Corpora

- A sample of a language
- Useful for studying the language
- Language is diverse
  - Big samples needed, to catch everything
  - Good tools needed, for large amounts of data
- Last 15 years
  - Big samples are easier to gather
  - Tools are better
  - Rapid growth in corpus methods

Web corpora

- Web is huge, free, easily accessible
- (Non-)linguists use it for lang. check/research
- Skewed?
  - Keller and Lapata 03:
    - web results match human judgements well
    - the large amount of data outweighs the “noise” problem
- Web importance as a resource is growing
  - David Crystal “Language and the Internet” 06:
    - “new linguistic medium that we cannot ignore”
- Web-corpus expertise is growing (WaCky etc.)

Steps to compile web corpora

(Sharoff, Baroni)

1. Get URL list for required language
   - ~500 most frequent word forms
     - not function words; for general-purpose corpora, words that do not belong to a spec. domain
   - 5000-6000 queries, 4 words, top 10 URLs
2. Download HTML pages
3. Normalize encoding (to UTF-8)
4. HTML clean-up
   - boilerplate removal: HTML tags, Java code, navigation frames,…
5. Extract meta-data (URL, title, date,…)
6. Linguistic annotation

Steps for JapWaC

- URL list of pages in Japanese provided by Serge Sharoff
  - word, lemma and PoS frequency lists for Japanese, c.f. http://corpus.leeds.ac.uk/list.html
- Files downloaded and cleaned with BootCat
  - by Marco Baroni and others from the WaCky project, c.f. http://wacky.sslmit.unibo.it/
- Segmented, tokenised, tagged with Chasen
- Translated Chasen tags to English
- Converted to Sketch Engine format and loaded

Example file

The file size is 7038669 kB, showing first 1 kB.

<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	
ブログデビュー	ブログデビュー		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>

Basic corpus statistics

- 49,554 URLs (i.e. HTML files)
- 16,072 sites (2 domains)
- 12,759,201 sentences (Chasen)
- 409,384,411 tokens (Chasen)
- 7.3 GB filesize

tokens/file:

- 8,263 Average
- 5,001 Median
- 3 Min
- 170,693 Max

URL statistics

(top ranking domains, sites and keywords)

corpus-jp-url-stats.xls

	A	B	C	D	E	F	G
1	% tokens	tokens	% files	files			
2	71.16%	291,299,194.00	70.46%	34,911.00	Domain .jp		
3	28.84%	118,085,211.00	29.54%	14,633.00	Domain .com		
4							
5	13.84%	56,649,811.00	12.20%	6,044.00	Keyword blog		
6	11.38%	46,590,191.00	6.91%	3,424.00	Keyword .go		
7	6.05%	24,774,335.00	4.78%	2,366.00	Keyword archives		
8	5.47%	22,398,942.00	5.96%	2,953.00	Keyword .or		
9	5.22%	21,362,799.00	5.51%	2,729.00	Keyword .ac		
10	4.13%	16,922,445.00	3.32%	1,646.00	Site blog.livedoor.jp		
11	4.10%	16,794,478.00	3.68%	1,821.00	Keyword diary		
12	3.58%	14,644,317.00	2.79%	1,380.00	Keyword .cocolog		
13	3.42%	14,009,671.00	2.61%	1,294.00	Keyword .exblog		
14	3.33%	13,647,702.00	1.39%	688.00	Keyword gijiroku		
15	3.14%	12,854,243.00	2.80%	1,385.00	Keyword .nifty		
16	3.12%	12,752,477.00	0.68%	335.00	Keyword .ndl		
17	3.06%	12,544,338.00	0.63%	311.00	Site kokkai.ndl.go.jp		
18	2.71%	11,082,044.00	2.41%	1,192.00	Keyword net		
19	2.55%	10,435,099.00	3.73%	1,850.00	Keyword news		
20	2.53%	10,338,210.00	1.96%	969.00	Keyword .geocities		

コマンド SCRL

9

Chasen POS statistics

corpus-jp-chasen-stats.xls

	A	B	C	D
1	Chasen tag	English	Tokens	% tokens
2	名詞-一般	N.g	50,121,542	12.24%
3	動詞-自立	V.free	32,746,478	8.00%
4	助詞-格助詞-一般	P.c.g	31,582,601	7.71%
5	助動詞	Aux	28,798,730	7.03%
6	記号-アルファベット	Sym.a	27,583,164	6.74%
7	名詞-サ変接続	N.Vs	21,704,856	5.30%
8	名詞-数	N.Num	20,160,119	4.92%
9	記号-読点	Sym.c	17,328,451	4.23%
10	助詞-連体化	P.prenom	14,786,763	3.61%
11	助詞-接続助詞	P.Conj	13,395,830	3.27%
12	記号-一般	Sym.g	13,312,393	3.25%
13	記号-句点	Sym.p	13,260,780	3.24%
14	助詞-係助詞	P.bind	12,854,037	3.14%
15	名詞-非自立-一般	N.bnd.g	7,582,296	1.85%
16	名詞-接尾-一般	N.Suff.g	6,945,415	1.70%
17	動詞-非自立	V.bnd	6,884,053	1.68%
18	*-知語	Unknown	6,857,818	1.68%
19	記号-括弧開	Sym.bc	6,171,426	1.51%
20	記号-括弧開	Sym.bo	6,061,836	1.48%
21	名詞-副詞可能	N.Adv	4,831,851	1.18%
22	名詞-接尾-助数詞	N.Suff.msr	4,543,767	1.11%
23	名詞-形容動詞語幹	N.Ana	4,436,970	1.08%
24	名詞-代名詞-一般	N.Pron.g	4,081,314	1.00%
25	形容詞-自立	Ai.free	3,871,216	0.95%
26	副詞-一般	Adv.g	3,603,890	0.88%
27	助詞-格助詞-引用	P.c.r	3,521,270	0.86%
28	助詞-格助詞-連語	P.c.Phr	3,514,071	0.86%
29	連体詞	Adn	3,178,853	0.78%

コマンド

10

The Sketch Engine

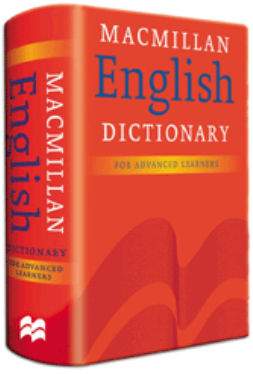
- Leading corpus query system
- Any corpus, ***any language***
- Web-based
  - *No software to install*
- Concordance
- **Word sketches**
  - "one-page, corpus based account of a word's grammatical and collocational behaviour"
- Thesaurus
- Word Sketch Difference

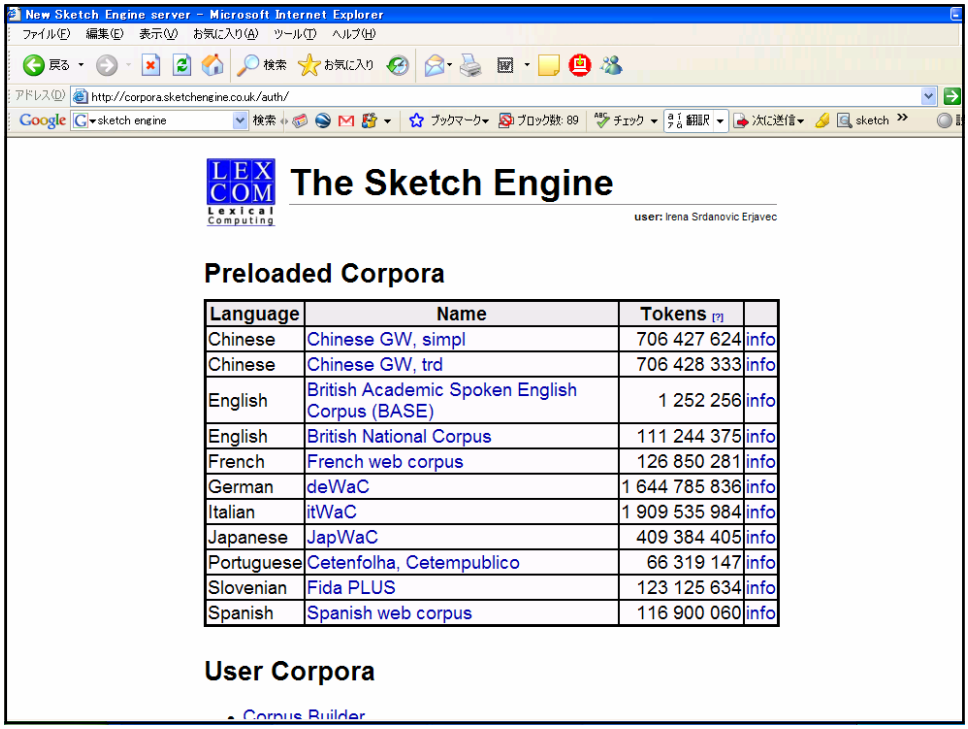
Use of Sketch Engine

- Lexicography
- Language learning
- Linguistic research

.
.
.

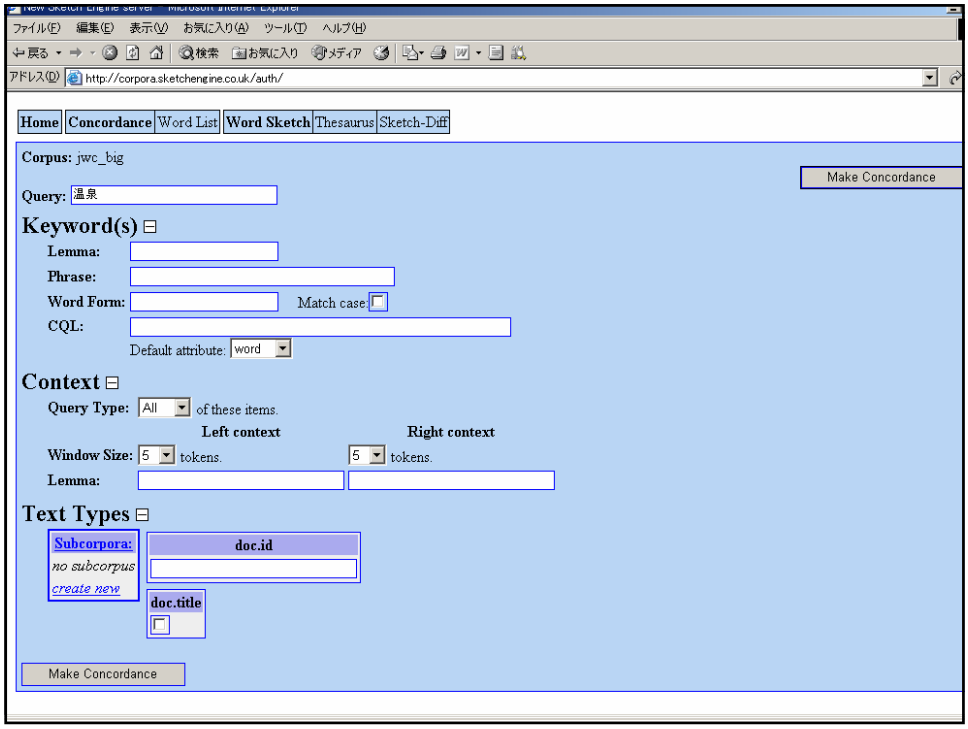
Macmillan English Dictionary
For Advanced Learners
Ed: Rundell, 2002

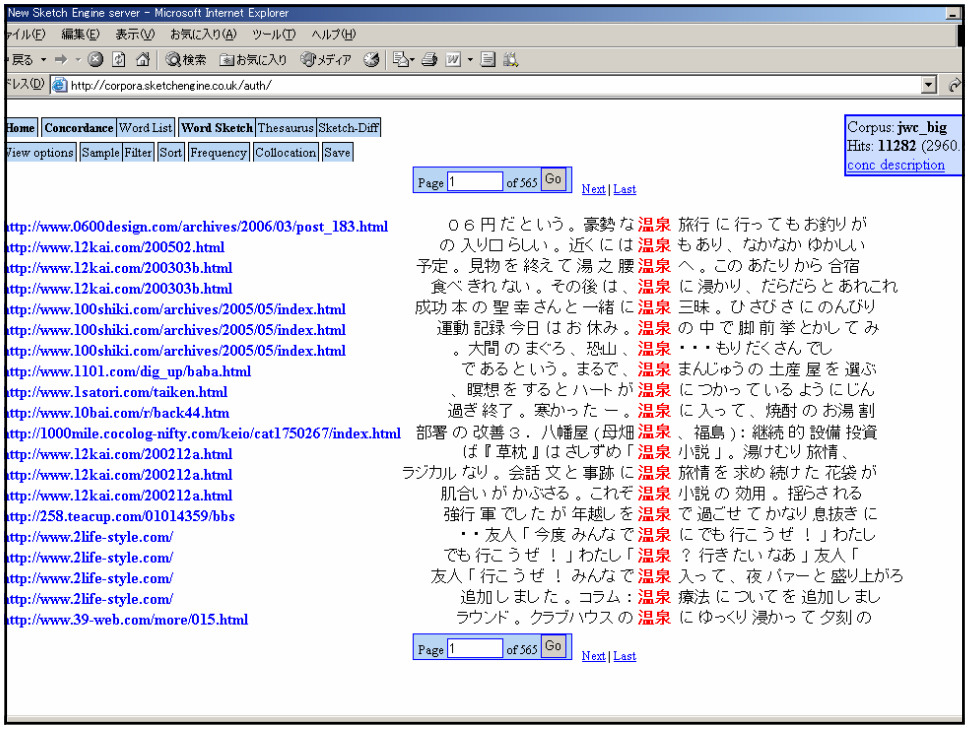


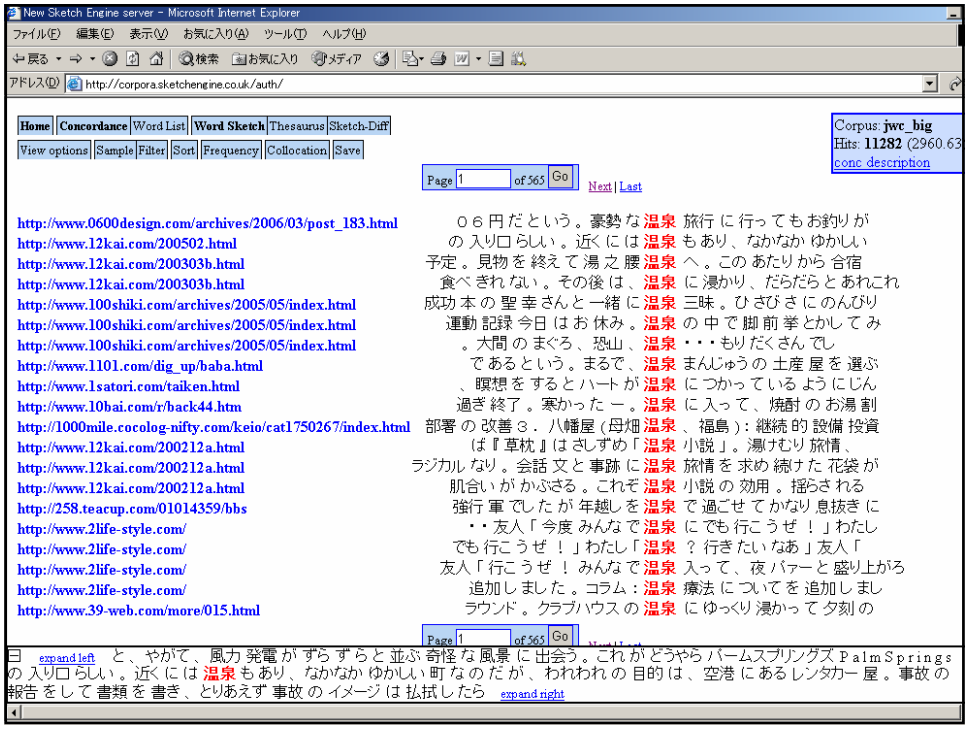


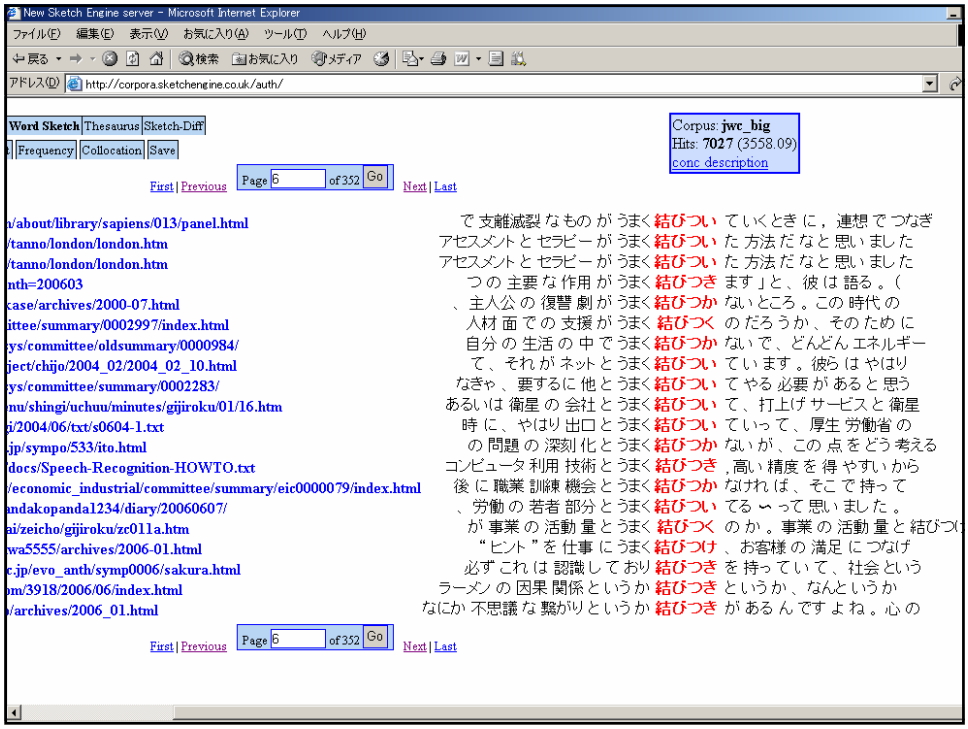
Creating SkE for Japanese

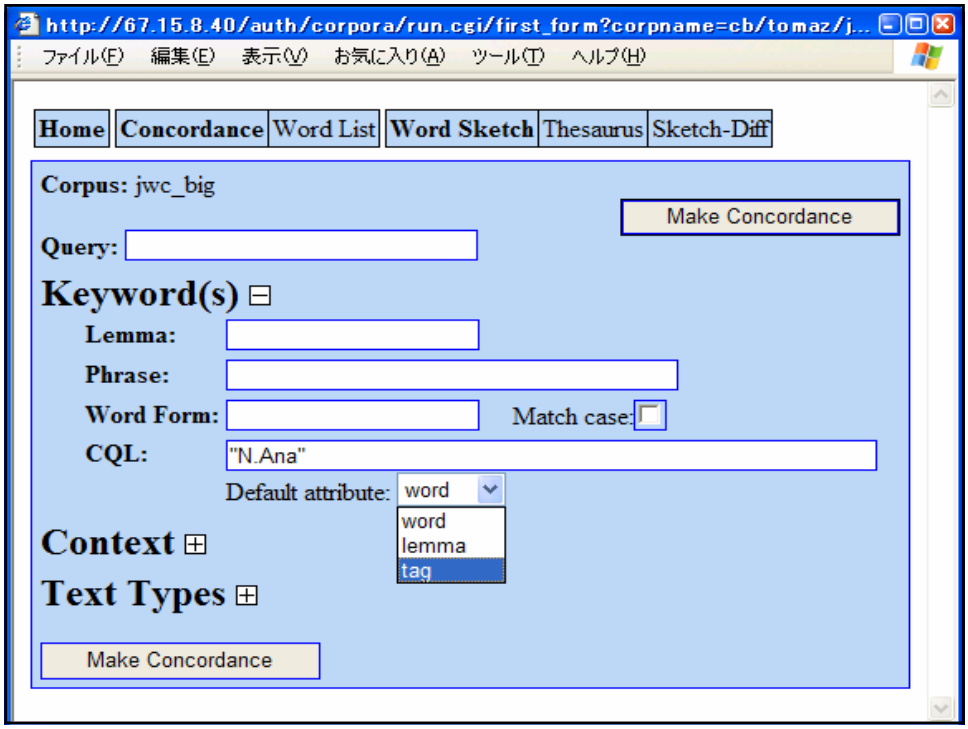
1. Load JapWaC into SkE
2. Write gram relations for Japanese
   - Chasen POS (as used for jaSlo)
3. Compile word sketches
4. Recompute scores in WS
5. Compile thesaurus

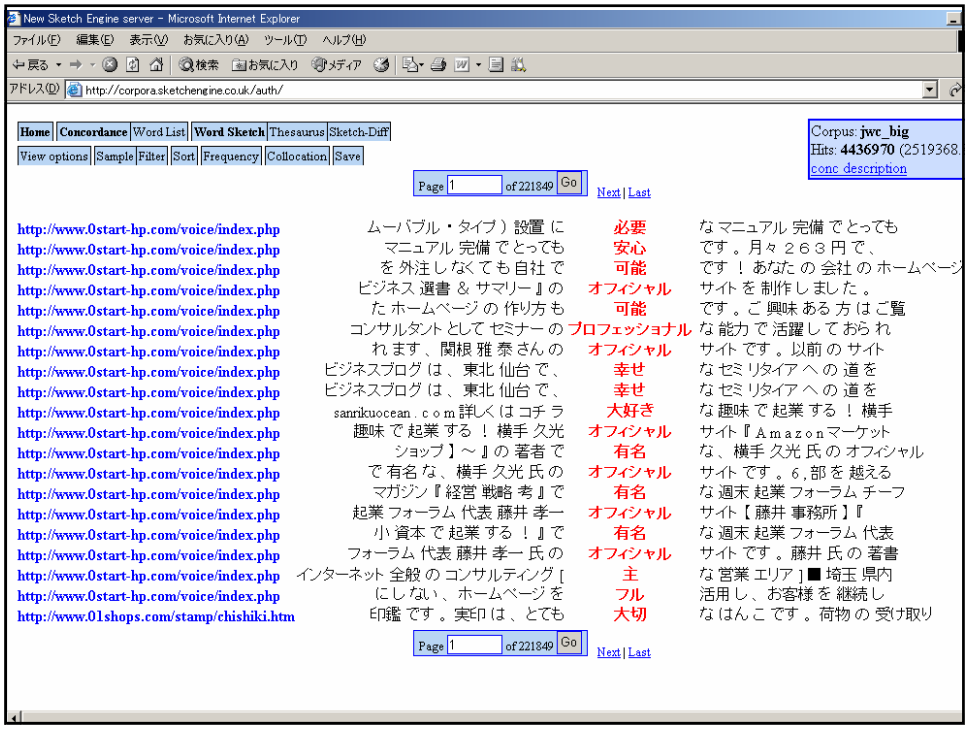

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り メディア
アドレス(D) http://corpora.sketchengine.co.uk/auth/
Home Concordance Word List Word Sketch Thesaurus Sketch-Diff
Corpus: jwc_big
Make Concordance
Query: 温泉
Keyword(s)
Lemma:
Phrase:
Word Form:
Match case:
CQL:
Default attribute: word
Context
Query Type: All of these items.
Left context
Right context
Window Size: 5 tokens. 5 tokens.
Lemma:
Text Types
Subcorpora:
no subcorpus
create new
doc.id
doc.title
Make Concordance

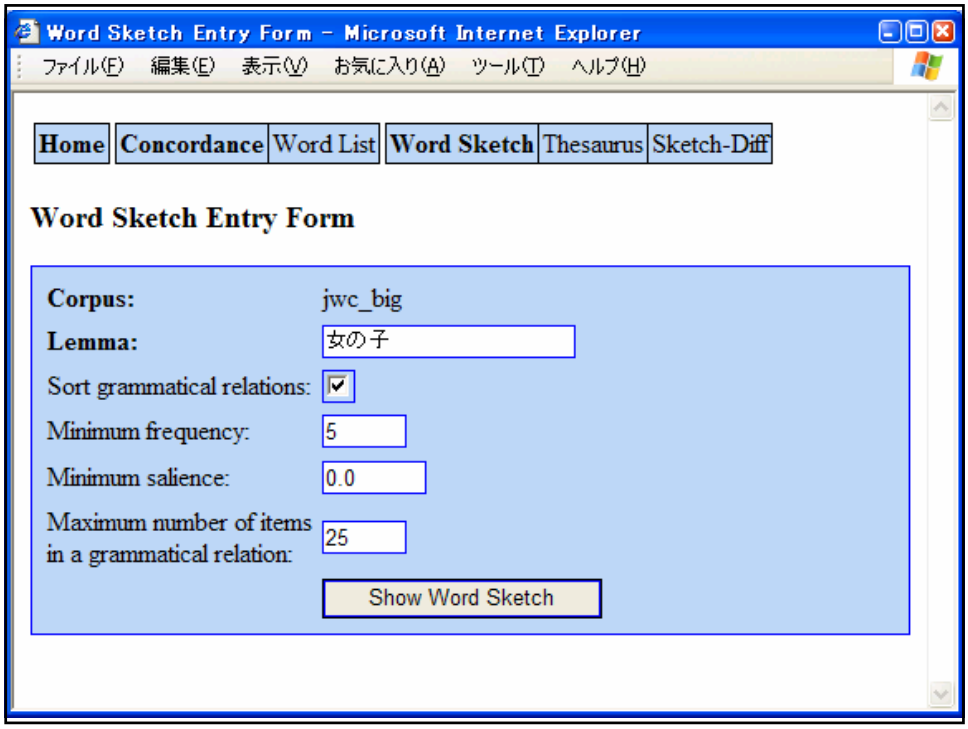

New Sketch Engine server - Microsoft Internet Explorer
http://corpora.sketchengine.co.uk/auth/
Home Concordance Word List Word Sketch Thesaurus Sketch-Diff
View options Sample Filter Sort Frequency Collocation Save
Corpus: jwc_big
Hits: 11282 (2960.
conc description
Page 1 of 565 Go
Next | Last
http://www.0600design.com/archives/2006/03/post_183.html ０６円だという。豪勢な温泉旅行に行ってもお釣りが
http://www.12kai.com/200502.html の入り口らしい。近くには温泉もあり、なかなかゆかしい
http://www.12kai.com/200303b.html 予定。見物を終えて湯之腰温泉へ。この あたりから合宿
http://www.12kai.com/200303b.html 食べきれない。その後は、温泉に浸かり、だらだらとあれこれ
http://www.100shiki.com/archives/2005/05/index.html 成功本の聖幸さんと一緒に温泉三昧。ひさびさにのんびり
http://www.100shiki.com/archives/2005/05/index.html 運動記録 今日はお休み。温泉の中で脚前挙とかしてみ
http://www.100shiki.com/archives/2005/05/index.html 。大間のまぐろ、恐山、温泉・・・もりだくさんでし
http://www.1101.com/dig_up/baba.html であるという。まるで、温泉まんじゅうの土産屋を選ぶ
http://www.1satori.com/taiken.html 、瞑想をするとハートが温泉につかっているようにじん
http://www.10bai.com/r/back44.htm 過ぎ終了。寒かった ー。温泉に入って、焼酎のお湯割
http://1000mile.cocolog-nifty.com/keio/cat1750267/index.html 部署の改善３．八幡屋(母畑温泉、福島)：継続的設備投資
http://www.12kai.com/200212a.html ば『草枕』はさしずめ「温泉小説」。湯けむり旅情、
http://www.12kai.com/200212a.html ラジカルなり。会話文と事跡に温泉旅情を求め続けた花袋が
http://www.12kai.com/200212a.html 肌合いがかぶさる。これぞ温泉小説の効用。揺らされる
http://258.teacup.com/01014359/bbs 強行軍でしたが年越しを温泉で過ごせてかなり息抜きに
http://www.2life-style.com/ ・・友人「今度みんなで温泉にでも行こうぜ！」わたし
http://www.2life-style.com/ でも行こうぜ！」わたし「温泉？行きたいなあ」友人「
http://www.2life-style.com/ 友人「行こうぜ！みんなで温泉入って、夜バァーと盛り上がろ
http://www.2life-style.com/ 追加しました。コラム：温泉療法についてを追加しまし
http://www.39-web.com/more/015.html ラウンド。クラブハウスの温泉にゆっくり浸かって夕刻の
Page 1 of 565 Go
Next | Last


New Sketch Engine server - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://corpora.sketchengine.co.uk/auth/
Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff
View options | Sample | Filter | Sort | Frequency | Collocation | Save
Corpus: jwc_big
Hits: 11282 (2960.63
conc description
Page 1 of 565 Go
Next | Last
http://www.0600design.com/archives/2006/03/post_183.html ０６円だという。豪勢な温泉旅行に行ってもお釣りが
http://www.12kai.com/200502.html の入り口らしい。近くには温泉もあり、なかなかゆかしい
http://www.12kai.com/200303b.html 予定。見物を終えて湯之腰温泉へ。この あたりから合宿
http://www.12kai.com/200303b.html 食べきれない。その後は、温泉に浸かり、だらだらとあれこれ
http://www.100shiki.com/archives/2005/05/index.html 成功本の聖幸さんと一緒に温泉三昧。ひさびさにのんびり
http://www.100shiki.com/archives/2005/05/index.html 運動記録 今日はお休み。温泉の中で脚前挙とかしてみ
http://www.100shiki.com/archives/2005/05/index.html 。大間のまぐろ、恐山、温泉・・・もりだくさんでし
http://www.1101.com/dig_up/baba.html であるという。まるで、温泉まんじゅうの土産屋を選ぶ
http://www.1satori.com/taiken.html 、瞑想をするとハートが温泉につかっているようにじん
http://www.10bai.com/r/back44.htm 過ぎ終了。寒かったー。温泉に入って、焼酎のお湯割
http://1000mile.cocolog-nifty.com/keio/cat1750267/index.html 部署の改善３．八幡屋(母畑温泉、福島):継続的設備投資
http://www.12kai.com/200212a.html ば『草枕』はさしずめ「温泉小説」。湯けむり旅情、
http://www.12kai.com/200212a.html ラジカルなり。会話文と事跡に温泉旅情を求め続けた花袋が
http://www.12kai.com/200212a.html 肌合いがかぶさる。これぞ温泉小説の効用。揺らされる
http://258.teacup.com/01014359/bbs 強行軍でしたが年越しを温泉で過ごせてかなり息抜きに
http://www.2life-style.com/ ・・友人「今度みんなで温泉にでも行こうぜ！」わたし
http://www.2life-style.com/ でも行こうぜ！」わたし「温泉？行きたいなあ」友人「
http://www.2life-style.com/ 友人「行こうぜ！みんなで温泉入って、夜バーと盛り上がろ
http://www.2life-style.com/ 追加しました。コラム：温泉療法についてを追加しまし
http://www.39-web.com/more/015.html ラウンド。クラブハウスの温泉にゆっくり浸かって夕刻の
Page 1 of 565 Go
日 expand left と、やがて、風力発電がずらずらと並ぶ奇怪な風景に出会う。これがどうやらパームスプリングズ PalmSprings の入り口らしい。近くには温泉もあり、なかなかゆかしい町なのだが、われわれの目的は、空港にあるレンタカー屋。事故の報告をして書類を書き、とりあえず事故のイメージは払拭したら expand right


New Sketch Engine server - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://corpora.sketchengine.co.uk/auth/
Word Sketch | Thesaurus | Sketch-Diff
Frequency | Collocation | Save
Corpus: jwc_big
Hits: 7027 (3558.09)
conc description
First | Previous Page 6 of 352 Go Next | Last
u/about/library/sapiens/013/panel.html で支離滅裂なものがうまく結びついていくときに，連想でつなぎ
/tanno/london/london.htm アセスメントとセラピーがうまく結びついた方法だなと思いました
/tanno/london/london.htm アセスメントとセラピーがうまく結びついた方法だなと思いました
nth=200603 つの主要な作用がうまく結びつきます」と、彼は語る。(
case/archives/2000-07.html 、主人公の復讐劇がうまく結びつかないところ。この時代の
ittee/summary/0002997/index.html 人材面での支援がうまく結びつくのだろうか、そのために
ys/committee/oldsummary/0000984/ 自分の生活の中でうまく結びつかないで、どんどんエネルギー
ject/chijo/2004_02/2004_02_10.html て、それがネットとうまく結びついています。彼らはやはり
ys/committee/summary/0002283/ なぎゃ、要するに他とうまく結びついてやる必要があると思う
nu/shingi/uchuu/minutes/gijiroku/01/16.htm あるいは衛星の会社とうまく結びついて、打上げサービスと衛星
j/2004/06/txt/s0604-1.txt 時に、やはり出口とうまく結びついていって、厚生労働省の
jp/sympo/533/ito.html の問題の深刻化とうまく結びつかないが、この点をどう考える
docs/Speech-Recognition-HOWTO.txt コンピュータ利用技術とうまく結びつき，高い精度を得やすいから
/economic_industrial/committee/summary/eic0000079/index.html 後に職業訓練機会とうまく結びつかなければ、そこで持って
ndakopanda1234/diary/20060607/ 、労働の若者部分とうまく結びついてるゃって思いました。
ai/zeicho/gijiroku/zc011a.htm が事業の活動量とうまく結びつくのか。事業の活動量と結びつ
wa5555/archives/2006-01.html “ヒント”を仕事にうまく結びつけ、お客様の満足につなげ
c.jp/evo_anth/symp0006/sakura.html 必ずこれは認識しており結びつきを持っていて、社会という
m/3918/2006/06/index.html ラーメンの因果関係というか結びつきというか、なんというか
/archives/2006_01.html なにか不思議な繋がりというか結びつきがあるんですよね。心の
First | Previous Page 6 of 352 Go Next | Last

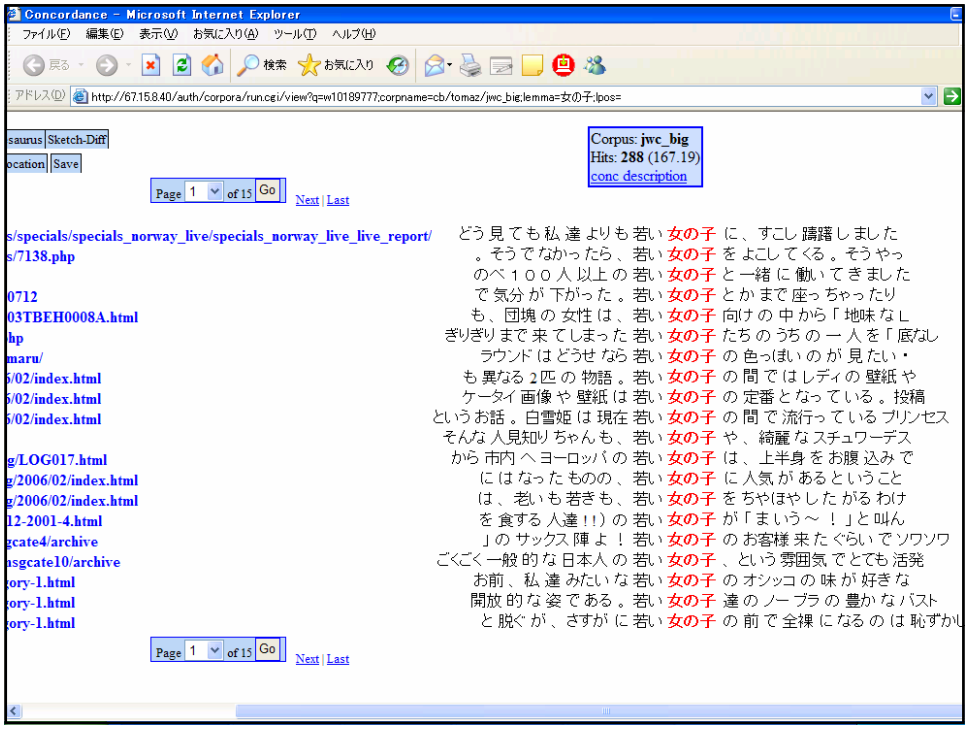

http://67.15.8.40/auth/corpora/run.cgi/first_form?corpname=cb/tomaz/j...
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
Home
Concordance
Word List
Word Sketch
Thesaurus
Sketch-Diff
Corpus: jwc_big
Make Concordance
Query:
Keyword(s)
Lemma:
Phrase:
Word Form:
Match case:
CQL:
"N.Ana"
Default attribute:
word
word
lemma
tag
Context
Text Types
Make Concordance


New Sketch Engine server - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://corpora.sketchengine.co.uk/auth/
Home
Concordance
Word List
Word Sketch
Thesaurus
Sketch-Diff
View options
Sample
Filter
Sort
Frequency
Collocation
Save
Corpus: jwc_big
Hits: 4436970 (2519368.
conc description
Page 1 of 221849 Go
Next | Last
http://www.0start-hp.com/voice/index.php ムーバブル・タイプ）設置 に 必要 な マニュアル 完備 で とっても
http://www.0start-hp.com/voice/index.php マニュアル 完備 で とっても 安心 です 。月々 ２６３ 円 で、
http://www.0start-hp.com/voice/index.php を 外注 し なく ても 自社 で 可能 です ！ あなた の 会社 の ホームページ
http://www.0start-hp.com/voice/index.php ビジネス 選書 ＆ サマリー 』の オフィシャル サイト を 制作 し ました 。
http://www.0start-hp.com/voice/index.php た ホームページ の 作り方 も 可能 です 。ご 興味 ある 方 は ご覧
http://www.0start-hp.com/voice/index.php コンサルタント として セミナー の プロフェッショナル な 能力 で 活躍 し て おら れ
http://www.0start-hp.com/voice/index.php れ ます 、関根 雅 泰 さん の オフィシャル サイト です 。以前 の サイト
http://www.0start-hp.com/voice/index.php ビジネスブログ は 、東北 仙台 で、 幸せ な セミリタイア へ の 道 を
http://www.0start-hp.com/voice/index.php ビジネスブログ は 、東北 仙台 で、 幸せ な セミリタイア へ の 道 を
http://www.0start-hp.com/voice/index.php sanrikuocean . c o m 詳しく は コチ ラ 大好き な 趣味 で 起業 する ！ 横手
http://www.0start-hp.com/voice/index.php 趣味 で 起業 する ！ 横手 久光 オフィシャル サイト 『 A m a z o n マーケット
http://www.0start-hp.com/voice/index.php ショップ 】～ 』の 著者 で 有名 な、横手 久光 氏 の オフィシャル
http://www.0start-hp.com/voice/index.php で 有名 な、横手 久光 氏 の オフィシャル サイト です 。6 , 部 を 越える
http://www.0start-hp.com/voice/index.php マガジン 『 経営 戦略 考 』で 有名 な 週末 起業 フォーラム チーフ
http://www.0start-hp.com/voice/index.php 起業 フォーラム 代表 藤井 孝一 オフィシャル サイト 【 藤井 事務所 】『
http://www.0start-hp.com/voice/index.php 小 資本 で 起業 する ！ 』で 有名 な 週末 起業 フォーラム 代表
http://www.0start-hp.com/voice/index.php フォーラム 代表 藤井 孝一 氏 の オフィシャル サイト です 。藤井 氏 の 著書
http://www.0start-hp.com/voice/index.php インターネット 全般 の コンサルティング [主 な 営業 エリア] ■ 埼玉 県内
http://www.0start-hp.com/voice/index.php にし ない 、ホームページ を フル 活用 し、お客様 を 継続 し
http://www.01shops.com/stamp/chishiki.htm 印鑑 です 。実印 は 、とても 大切 な はんこ です 。荷物 の 受け取り
Page 1 of 221849 Go
Next | Last



Word Sketch examples

- WS for 女の子 (noun)
- WS for 冷たい (adjective)
- WS for 書く (verb)



Thesaurus, WS Diff example(1)

女の子 jwc_big freq = 16309

男の子	0.463	少女 0.389 少年 0.362 青年 0.289 妹 0.283 赤ちゃん 0.28 おじさん 0.278 老人 0.268 誰か 0.25
男性	0.403	子 0.391 女性 0.389 女 0.38 娘 0.38 子ども 0.363 子供 0.359 友達 0.352 男 0.337 息子 0.329 彼女 0.324 生徒 0.323 お母さん 0.322 若者 0.321 人達 0.319 友人 0.314 学生 0.306 母親 0.301 日本人 0.296 奥さん 0.291 俺 0.285 夫 0.279 母 0.279 僕 0.27 ママ 0.269 ぼく 0.268 妻 0.266 仲間 0.262 わたし 0.261 家族 0.256 あなた 0.255 先生 0.254 お父さん 0.251 父親 0.251 教師 0.248 親 0.248

- WS Diff for 女の子 and 男の子

Thesaurus, WS Diff example (2)

冷たい jwc_big freq = 7084

暖かい	0.384	温かい 0.364 熱い 0.323 優しい 0.299 やさしい 0.246 心地よい 0.222 あたたかい 0.212 あったかい 0.208 気持ちよい 0.201
寒い	0.319	涼しい 0.279 辛い 0.264 寂しい 0.241 痛い 0.238 怖い 0.236 暑い 0.235 汚い 0.231 恐ろしい 0.224 珍しい 0.221 悲しい 0.209 忙しい 0.208 まずい 0.207 懐かしい 0.207 いや 0.204 嫌 0.203 つらい 0.201

- WS Diff for 寒い and 冷たい

「温泉」 example

- WS for 温泉

Future work

- More metadata in the corpus:
  - date, title, author; text typology
- More data cleaning
- Japanese corpus for HLT research:
  - sampling only 10 consecutive sentences, 100M
  - would be available for download with Creative Commons license
- For native speakers' and learners' use:
  - original Chasen tags, Chasen kana
  - Ruby romaji, furigana in examples
- Connecting to jaSlo, Natsume system
- More advanced relations (MWU etc.), Cabocha?
- Load other corpora into SkE (Kotonoha, AB)

Access to JapWaC & SkE

- *http://www.sketchengine.co.uk*
- Free 30-day trial
- Self-registration
- Japanese, Chinese, English, French, German, Italian, Spanish, Portuguese, Slovene
- Also gives access to WebBootCaT
  - "instant web corpora"

Thank you for your attention!