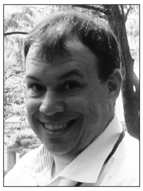


Edo Forsythe

In this column, we explore the issue of teachers and technology—not just as it relates to CALL solutions, but also to Internet, software, and hardware concerns that all teachers face. We invite readers to submit articles on their areas of interest. Please contact the editor before submitting.

Email: tlt-wired@jalt-publications.org • Web: http://jalt-publications.org/tlt/departments/tlt-wired

Using the Sketch Engine Corpus Query Tool for Language Teaching

Keith Barrs
Hiroshima Shudo University
barrs@shudo-u.ac.jp

Corpora, and the tools used to query them, have seen rapid advancements in the last decade. These advancements have primarily been concerned with the exploitation of web-derived linguistic data through the development of sophisticated tools that can handle the crawling, processing, and analysing of the data extracted from the World Wide Web. These web corpora, often made up of billions of words, rather than the millions associated with more traditional corpora like the British National Corpus (BNC), can be loaded into web-based corpus query systems, such as Sketch Engine, CQPweb, and Colibri2, allowing corpus querying to be done anywhere. This article gives a brief overview of the rise of web corpora and web-based corpus query systems, and outlines some of the ways in which the Sketch Engine—a web-based corpus query system built around a huge number of web-derived corpora—can be exploited for educational purposes.

The Rise of Web Corpora and Web-Based Corpus Query Systems

The benefits of exploiting web text within corpus linguistics were recognised over 15 years ago, with Kilgarriff (2001) stating that “the web is with us, giving access to colossal quantities of text, of any number of varieties, at the click of a button, for free” (p. 344). Bernardini, Baroni, and Evert (2006) put forward four different conceptualisations of the relationship between the web text and corpus linguistics: (a) web as a corpus shop, whereby researchers download texts retrieved by search engines and make disposable corpora; (b) web as a corpus surrogate, whereby the web is accessed through search engines to achieve tasks such as translations; (c) web as a corpus proper, which looks at the web as a whole in terms of web English; and (d) a mega-corpus mini-web, which involves combining large amounts of web-derived texts with corpus characteristics, such as part-of-speech annotation (pp. 10-14).

Going hand-in-hand with the rise of web corpora have been rapid advancements in the tools used to query them. The most groundbreaking advancements have been with the development of web-based corpus query systems, meaning that users can engage in corpus linguistics without needing to download the corpora or the software programs. For web corpora, this has been an essential development as it means users can easily and cheaply access ultra-large-scale corpora that would be too large and costly to distribute for download. This has untied corpus linguistics from private computers and moved it into a cloud-based, mobile environment. Indeed, it is now possible to make a rapid query of a ten-billion word corpus of any of the world’s major languages simply through using an Internet connected device.

The Sketch Engine

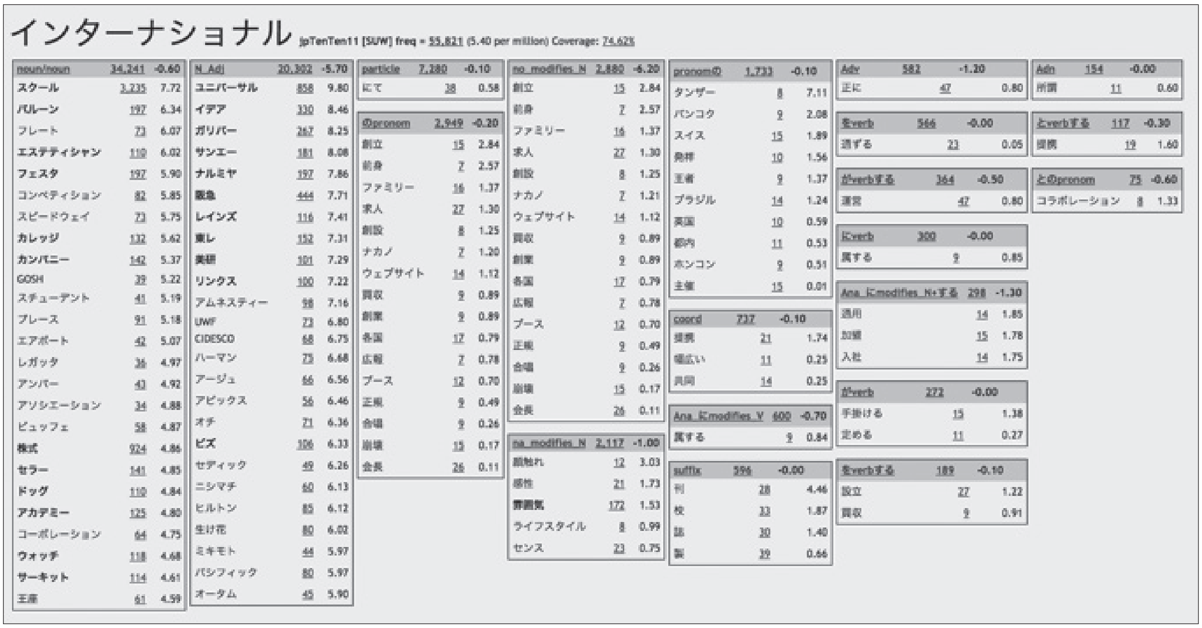
One of the first and most widely used platforms to bring together web corpora and web-based corpus query tools is the Sketch Engine <https://www.sketchengine.co.uk>. This is an online corpus interface that houses over 200 corpora of over 80 languages. At its core it includes a family of 31 web corpora of most of the world’s major languages, called the TenTen family. Each one contains between two and 15 billion words that have been crawled and processed in a similar manner. Whilst the corpora are not balanced in the traditional sense of balanced corpora, such as the Brown Corpus on written American English and the LOB Corpus on written British English, the fact that they are part of a family makes it possible to compare the behavior of words between the different web-based languages.

The central function of the Sketch Engine is the *word sketch*, which is used to reveal how words behave collocationally and grammatically within a corpus. These were originally developed in order to help with lexicography, and are now used widely within language research and education in general. Figure 1 shows a word sketch from the jpTenTen11 corpus of Japanese web text for the English loanword インターナショナル (*intaanashonaru*) [international]. Through the word sketch, a rich overview of the behaviour of this word in all of its 55,821 instances in the corpus becomes quickly and easily visible, and is something that would be all but impossible through introspection alone. The word can be further investigated through analysing the concordance lines, and comparing the loanword with its native, near-semantic equivalent of 国際 (*kokusai)* [international]. This is possible using the innovative function of the *sketch-diff*, which allows the user to compare the collocational and grammatical behaviour of two similar words to draw out the important differences in usage. This word sketch could then be compared with the English language form of *international*, generated from the enTenTen12 corpus.

The Sketch Engine for Language Learning

Word sketches can be brought into the classroom through access to the full Sketch Engine system, via individual or institutional licenses after a one-month free trial. This allows full access to a wide variety of corpora in a large number of languages, and it allows the student to exploit a range of corpus query tools for activities such as concordancing, producing word frequency lists, and comparing near synonyms. There has been a recent, pedagogically-focused development of the Sketch Engine with the introduction of the Sketch Engine for Language Learning (SkELL) <https://skell.sketchengine.co.uk/run.cgi/skell>, which is a free, online, stripped-down version of the full software program.

SkELL references a one-billion word web corpus of English, with a simple, student-friendly user interface. It is built around three primary functions: concordancing, word sketches, and a thesaurus. The concordancing tool returns 40 corpus-based examples of the keyword in context. These can be used to develop reading skills such as context clues, where students try to define a word by examining its linguistic context. Using copy and paste, the concordance lines can also be used to create a fill-in-the-gap exercise (see Appendix) where the search word, or words in the context, can be blanked out for the students to fill in. For more collocational and grammatical detail, the word sketch function of SkELL produces compact versions of the full word sketches in the Sketch Engine, and can be used to get a quick overview of the different senses of a word. Figure 2 shows that the collocates are arranged by grammatical category, which means focus can be given to specific collocations within particular grammatical structures and categories, allowing a rich understanding of the word's behaviour. Furthermore,



インターナショナル jpTenTen11 [SUW] freq = 55,821 (5.40 per million) Coverage: 74.62%

noun/noun	34,241	-0.60
スクール	3,235	7.72
バルーン	197	6.34
プレート	73	6.07
エステティシャン	110	6.02
フェスタ	197	5.90
コンペティション	82	5.85
スピードウェイ	73	5.75
カレッジ	132	5.62
カンパニー	142	5.37
GOSH	39	5.22
スチューデント	41	5.19
プレース	91	5.18
エアポート	42	5.07
レガッタ	36	4.97
アンバー	43	4.92
アソシエーション	34	4.88
ビュッフェ	58	4.87
株式	924	4.86
セラー	141	4.85
ドッグ	110	4.84
アカデミー	125	4.80
コーポレーション	64	4.75
ウォッチ	118	4.68
サーキット	114	4.61
王座	61	4.59

N_Adj	20,302	-5.70
ユニバーサル	858	9.80
イデア	330	8.46
ガリバー	267	8.25
サンエー	181	8.08
ナルミヤ	197	7.86
阪急	444	7.71
レインズ	116	7.41
東レ	152	7.31
美研	101	7.29
リンクス	100	7.22
アムネスティー	98	7.16
UWF	73	6.80
CIDESCO	68	6.75
ハーマン	73	6.68
アージュ	66	6.56
アピックス	56	6.46
オチ	71	6.36
ビズ	106	6.33
セディック	49	6.26
ニシマチ	60	6.13
ヒルトン	85	6.12
生け花	80	6.02
ミキモト	44	5.97
パシフィック	80	5.97
オータム	45	5.90

particle	7,280	-0.10
にて	38	0.58

のpronom	2,949	-0.20
創立	15	2.84
前身	7	2.57
ファミリー	16	1.37
求人	27	1.30
創設	8	1.25
ナカノ	7	1.20
ウェブサイト	14	1.12
買収	9	0.89
創業	9	0.89
各国	17	0.79
広報	7	0.78
ブース	12	0.70
正規	9	0.49
合唱	9	0.26
崩壊	15	0.17
会長	26	0.11

no_modifies_N	2,880	-6.20
創立	15	2.84
前身	7	2.57
ファミリー	16	1.37
求人	27	1.30
創設	8	1.25
ナカノ	7	1.21
ウェブサイト	14	1.12
買収	9	0.89
創業	9	0.89
各国	17	0.79
広報	7	0.78
ブース	12	0.70
正規	9	0.49
合唱	9	0.26
崩壊	15	0.17
会長	26	0.11

na_modifies_N	2,117	-1.00
顔触れ	12	3.03
感性	21	1.73
雰囲気	172	1.53
ライフスタイル	8	0.99
センス	23	0.75

pronomの	1,733	-0.10
タンザー	8	7.11
バンコク	9	2.08
スイス	15	1.89
発祥	10	1.56
王者	9	1.37
ブラジル	14	1.24
英国	10	0.59
都内	11	0.53
ホンコン	9	0.51
主催	15	0.01

coord	737	-0.10
提携	21	1.74
幅広い	11	0.25
共同	14	0.25

Ana_にmodifies_V	600	-0.70
属する	9	0.84

suffix	596	-0.00
判	28	4.46
校	33	1.87
語	30	1.40
製	39	0.66

Adv	582	-1.20
正に	47	0.80

をverb	566	-0.00
通ずる	23	0.05

がverbする	364	-0.50
運営	47	0.80

にverb	300	-0.00
属する	9	0.85

Ana_にmodifies_N+する	298	-1.30
通用	14	1.85
加盟	15	1.78
入社	14	1.75

がverb	272	-0.00
手掛ける	15	1.38
定める	11	0.27

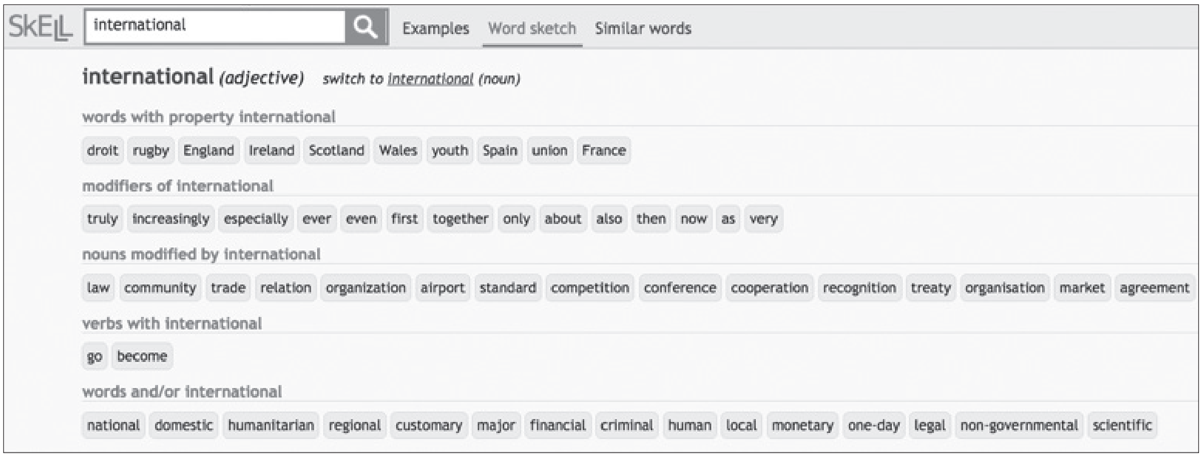
をverbする	189	-0.10
設立	27	1.22
買収	9	0.91

Adn	154	-0.00
所謂	11	0.60

とverbする	117	-0.30
提携	19	1.60

とのpronom	75	-0.60
コラボレーション	8	1.33

Figure 1. **A word sketch from the jpTenTen11 (Japanese) corpus for the English loanword インターナショナル (*intaanashonaru*) [international].**



international *(adjective)* *switch to international (noun)*

words with property international

droit rugby England Ireland Scotland Wales youth Spain union France

modifiers of international

truly increasingly especially ever even first together only about also then now as very

nouns modified by international

law community trade relation organization airport standard competition conference cooperation recognition treaty organisation market agreement

verbs with international

go become

words and/or international

national domestic humanitarian regional customary major financial criminal human local monetary one-day legal non-governmental scientific

Figure 2**. A word sketch produced in SkELL for the English word *****international*****.**

each of the collocates is clickable, which takes the user to new concordance lines of the search word in context with the selected collocation. The third function of producing lists of similar words returns around 20 synonyms of the search word, and can help with developing skills such as selecting alternative expressions for a writing task.

The most rewarding way in which corpus tools are put into practice in my classes is by encouraging students to pursue corpus-based analyses of the English language for their graduation theses. A particularly insightful study compared the meanings of Japanese loanwords in English with their counterparts in the Japanese language via a comparative analysis of their most frequent collocations. This was done by using the collocation function of the Sketch Engine to generate lists of the collocations, as well as by generating lists of similar words to the main search word using the thesaurus function. Another student used the tools within the Sketch Engine to create their own mini-corpus of Japan-related English web pages, and looked at the ways in which Japanese culture is described in English. Another study focused on English in the Japanese linguistic landscape, and used English corpora to check the frequencies and meanings of words that are used on Japanese shop signs.

Conclusion and Further Reading

Corpus linguistics as a method of investigating language has been greatly invigorated by the increasing availability of corpora and corpus query tools. This has been made possible through the rapid advancements made with web corpora, which allow them to be compiled with significantly less human and financial resources than has been previously possible. This, in turn, has led many of the corpora and corpus querying tools to become cloud-based, such as the Sketch Engine and SkELL tools, meaning that access to corpus linguistics has been opened up to anyone with Internet access who wants to investigate how language behaves. For language education in particular, the advancements in simplifying the access to large-scale corpora means that corpus-based data can become a regular part of language teaching, learning, and research. More details on using the Sketch Engine for learning English, including a large number of practical classroom exercises and examples, can be found in *Discovering English with the Sketch Engine* (Thomas, 2015).

References

Bernardini, S., Baroni, M., & Evert, S. (2006). A WaCky introduction. In M. Baroni & S. Bernardini (Eds.), *Wacky! Working Papers on the Web as Corpus* (pp. 9–40). Bologna, Italy: GEDIT.

Kilgarriff, A. (2001). Web as corpus. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 342–344). Lancaster, U.K.: University of Lancaster.

Thomas, J. (2015). *Discovering English with the Sketch Engine*. Brno: Versatile.

Editor's Note: The web has brought a myriad of tools to our students' fingertips, and Keith Barrs has shared an engaging tool for investigating how language is evolving and being used today. The appendix is available below. You can learn about many more tools and technology practices at JALTCALL 2016 this June at Tamagawa University. Registration begins soon, and details are available at <http://conference.jaltcall.org>. Together we'll learn how to keep our classes *Wired!*

Appendix

Context Clues Concordance Jigsaw Activity

Instructions (based on a class of 16 students)

- Make four groups. Give each group one of the four sets of 10 concordance lines.
- Ask groups to think about the missing word in their set (the missing word is the same for all the concordance lines in that set).
- Guide them to use the context around the missing words in the concordance lines to help make their decision. (For example, using background knowledge, part-of-speech, and meaning)
- Ask students to make notes of anything interesting that appeared in the concordance lines, for example multiple meanings of the word, figurative language, and spelling/grammatical mistakes.
- Reorganise the groups so there is one student representative from each of the four concordance sets in each of the four new groups.
- Give each group some extra copies of the concordance sets. Have the student representative for each group encourage the others to guess the missing word for their set, and then explain anything interesting that they had noticed about the concordance set in their previous group.
- For homework, ask students to make their own concordance set, with the keyword blanked out, for a vocabulary word related to the theme of the course.

Concordance Set 1

1	The preceding & following _______________s were very snowy.
2	Ice hockey is official national _______________ sport.
3	France has neither _______________ nor summer nor morals.
4	It looks like _______________ has finally arrived.
5	A late _______________ means a late spring.
6	The rice barrel was left empty during the ___________ months.
7	This course is offered during _______________ quarter only.
8	During _______________ temperatures frequently drop below freezing at night.
9	The area enjoys warm summers and mild _______________s.
10	_______________ storms swiftly obliterated his expensive engineering structures.

Concordance Set 2

1	This happy mood lasted roughly until last _______________.
2	The _______________ months were slightly above average.
3	The promotional activity last _______________ was rather less visibly aggressive.
4	Fall is here and colorful _______________ leaves are abundant.
5	_______________ has occurred every year since records began.
6	The _______________ sun was beginning to set.
7	_______________ and winter months are usually windy.
8	Spring and _______________ bring fairly mild weather.
9	But summer passed away and _______________ came.
10	The weather is beautiful–a lovely _______________ morning.

Concordance Set 3

1	Its expected completion date is _______________ 2015.
2	The _______________ hunting season may impact male mortality rates.
3	A full size box _______________ is 53 inches long.
4	The housing turnaround hit full stride last _______________.
5	The _______________ game is where coaches build goodwill.
6	A late winter means a late _______________.
7	Some students enjoyed sandy beaches on _______________ break.
8	The best bird watching times are late fall and early _______________.
9	New luxury hotels are _______________ing up everywhere.
10	The vast boundless valley colored by bright _______________ flowers.

Concordance Set 4

1	Remember fire bans apply during _______________ months.
2	My best summer story happened every _______________ night.
3	Average _______________ temperatures are around 22 degrees.
4	His black bear sounds were recorded right here last _______________.
5	The anticipated release date is _______________ 2013.
6	The _______________ was going–was gone.
7	Around every corner is another interesting _______________ home.
8	Extra buses are added during the high _______________ season.
9	There are _______________ festivals and farmers markets.
10	The days of _______________ heat defy description.

Answers

- Concordance set 1: Winter.
- Concordance set 2: Autumn.
- Concordance set 3: Spring.
- Concordance set 4: Summer.