

The Sketch Engine: ten years on

Adam Kilgarriff · Vít Baisa · Jan Bušta · Miloš Jakubíček ·
Vojtěch Kovář · Jan Michelfeit · Pavel Rychlý · Vít Suchomel

Published online: 10 July 2014
© Springer Berlin Heidelberg 2014

Abstract The Sketch Engine is a leading corpus tool, widely used in lexicography. Now, at 10 years old, it is mature software. The Sketch Engine website offers many ready-to-use corpora, and tools for users to build, upload and install their own corpora. The paper describes the core functions (word sketches, concordancing, thesaurus). It outlines the different kinds of users, and the approach taken to working with many different languages. It then reviews the kinds of corpora available in the Sketch Engine, gives a brief tour of some of the innovations from the last few years, and surveys other corpus tools and websites.

Keywords Corpora · Corpus lexicography · Corpus tools · Word sketches · Sketch Engine

1 Introduction

The Sketch Engine is a leading corpus tool. It has been widely used in lexicography. It is now 10 years since its launch (Kilgarriff et al. 2004).

Those 10 years have seen dramatic changes. They have seen the near-death of dictionaries on paper, at the hands of electronic dictionaries.¹ They have seen the

¹ The change is often lamented. Rundell (2012) celebrates it.

Communicated by Yukio Tono.

A. Kilgarriff (✉) · V. Baisa · J. Bušta · M. Jakubíček · V. Kovář · J. Michelfeit ·
P. Rychlý · V. Suchomel
Lexical Computing Ltd, Brighton, UK
e-mail: adam.kilgarriff@gmail.com

V. Baisa · J. Bušta · M. Jakubíček · V. Kovář · J. Michelfeit ·
P. Rychlý · V. Suchomel
Masaryk University, Brno, Czech Republic

emergence of entire new ecosystems of dictionaries on the web, with many new players (Google, weblio.jp, dictionary.com, Leo, Wordnik.com). Previously, the dominant players had been around for decades, even centuries—Longman (who published Johnson’s dictionary in 1754), Kenkyusha, OUP, Le Robert, Duden, Merriam-Webster.

In the world at large, we have seen the invention and world takeover of the smartphone. 1994–2004 saw the switch of most dictionary lookups from paper to electronic: 2004–2014 has seen them nearly all (in percentage terms) switch from computer to phone. (Just think how often your students look up words on their phones, versus how often they look them up in any other way.) Dictionaries are far, far more available and accessible than they were. The sheer number of dictionary lookups has risen many times over (even as—bitter irony—many dictionary companies have seen their income collapse).

This is all at the publishing end of the dictionary business. What about the lexicography end? Here, we have seen the corpus revolution (Hanks 2012). It started in Northern Europe in the 1980s and 1990s, and has been spreading. For Chinese, a first thoroughly corpus-based dictionary was probably Huang et al. (1997)’s classifier-noun collocation dictionary. For Arabic, it is Oxford University Press’s Oxford Arabic Dictionary (Arts 2014),² though this was not produced in Asia. In Japan, corpus lexicography started in bilingual dictionary projects such as the *WISDOM English-Japanese Dictionary* (Sanseido 2003, 2007), but a truly corpus-based monolingual dictionary of Japanese is yet to appear.

Thus the 10 years of the Sketch Engine have also been the 10 years of bringing corpora into Asian lexicography. The paper is a perspective on those changes.

In this paper we review

- the tool
- its users
- the languages covered
- the corpora accessible in it, and
- developments in the software over the past decade.

We finish by reviewing related work: other corpora, corpus websites and corpus tools as available for lexicography and corpus linguistics.

‘Sketch Engine’ refers to two different things: the software, and the web service. The web service includes, as well as the core software, a large number of corpora pre-loaded and ‘ready for use’, and tools for creating, installing and managing your own corpora. The paper covers both, with Sects. 2 and 6 focussing on the software, 3, 4, and 5, the web service.

² The Oxford Arabic Dictionary is a bilingual Arabic-English, English-Arabic dictionary. For bilingual dictionaries, a corpus is most relevant for the analysis of the source language. In this dictionary the source side of the Arabic-English half is the first corpus-based, dictionary-scale analysis of the Arabic lexicon.

catch (verb) Alternative PoS: **noun** (99356)
enTenTen12 freq = 1311175 (101.1 per million)

| object | 633,096 | 0.6 | subject | 222,952 | 0.3 | and/or | 29,216 | 0.0 | part_intrans | 233,411 | 5.4 | pro_object | 103,679 | 0.8 |
|-----------|---------|------|-----------|---------|------|-----------|--------|------|--------------|---------|------|------------|---------|------|
| glimpse | 17,529 | 9.34 | angler | 914 | 6.5 | prosecute | 301 | 6.32 | up | 219,753 | 8.74 | myself | 4,235 | 6.97 |
| fish | 26,116 | 8.55 | eye | 12,536 | 6.31 | punish | 428 | 6.15 | on | 7,069 | 7.81 | him | 17,863 | 6.58 |
| attention | 37,300 | 8.29 | fisherman | 860 | 6.15 | overheat | 93 | 5.65 | out | 5,814 | 3.48 | her | 6,422 | 6.22 |
| breath | 11,696 | 8.18 | surprise | 1,231 | 5.42 | convict | 268 | 5.46 | along | 156 | 2.44 | herself | 774 | 5.86 |
| eye | 40,725 | 7.9 | touchdown | 425 | 5.31 | bowl | 137 | 5.46 | away | 143 | 1.89 | himself | 1,520 | 5.55 |
| pass | 7,152 | 7.64 | Anyone | 468 | 5.05 | farm | 92 | 5.04 | through | 55 | 1.44 | me | 14,392 | 5.51 |
| sight | 8,250 | 7.54 | police | 1,615 | 4.8 | jail | 62 | 5.03 | upon | 28 | 1.18 | them | 18,263 | 5.48 |
| fire | 12,436 | 7.39 | 'll | 574 | 4.43 | imprison | 99 | 5.0 | | | | yourself | 2,300 | 5.4 |
| trout | 3,680 | 7.28 | breath | 559 | 4.33 | tangle | 47 | 4.82 | | | | ourselves | 326 | 4.44 |
| bus | 5,779 | 6.89 | fish | 1,093 | 4.26 | tame | 38 | 4.74 | | | | us | 3,912 | 4.26 |
| bass | 3,681 | 6.88 | TD | 172 | 4.2 | throw | 1,200 | 4.73 | | | | hers | 54 | 3.67 |
| ball | 8,172 | 6.83 | fisher | 141 | 4.15 | snag | 36 | 4.64 | | | | it | 22,088 | 3.12 |

| pp_in-i | 65,284 | 0.4 | pp_on-i | 22,933 | 0.3 | pp_by-i | 17,122 | 0.4 | pp_with-i | 9,683 | 0.1 | np_adj_comp | 6,638 | 0.3 |
|------------|--------|------|-------------|--------|------|-----------|--------|------|------------|-------|------|---------------|-------|-------|
| crossfire | 1,731 | 9.63 | videotape | 146 | 6.88 | surprise | 1,244 | 5.79 | pant | 821 | 5.77 | red-handed | 385 | 10.41 |
| trap | 2,058 | 7.98 | tape | 1,456 | 6.17 | fisherman | 274 | 5.56 | hooker | 24 | 5.01 | unprepared | 276 | 7.78 |
| middle | 3,334 | 6.89 | CCTV | 108 | 5.74 | paparazzi | 50 | 5.54 | contraband | 14 | 4.74 | flatfooted | 28 | 7.02 |
| headlight | 521 | 6.89 | wildfire | 101 | 5.71 | angler | 168 | 5.47 | tnt | 57 | 4.46 | unaware | 370 | 6.08 |
| adultery | 380 | 6.83 | fire | 2,889 | 5.7 | IR | 72 | 5.33 | prostitute | 48 | 4.2 | redhanded | 13 | 5.98 |
| cross-fire | 192 | 6.58 | jig | 70 | 5.07 | trawler | 27 | 5.04 | seine | 7 | 4.11 | one-handed | 19 | 5.57 |
| snare | 219 | 6.52 | theleupapul | 23 | 5.04 | Akismet | 20 | 5.02 | alcohol | 466 | 4.01 | off-balance | 15 | 5.48 |
| throat | 1,077 | 6.24 | minnow | 40 | 5.01 | fisher | 44 | 4.74 | baggle | 7 | 3.93 | unguarded | 14 | 5.2 |
| downpour | 190 | 6.18 | camera | 2,215 | 4.92 | avalanche | 41 | 4.42 | britches | 5 | 3.66 | single-handed | 5 | 4.14 |
| traps | 374 | 6.0 | bait | 176 | 4.9 | fielder | 22 | 4.13 | rave | 8 | 3.36 | unready | 4 | 4.08 |
| crosshairs | 121 | 5.8 | lure | 109 | 4.87 | catcher | 38 | 4.04 | powerstam | 3 | 3.33 | bare-handed | 3 | 3.84 |
| avalanche | 188 | 5.8 | wiretap | 27 | 4.79 | peloton | 11 | 4.03 | chaff | 6 | 3.31 | short-handed | 4 | 3.75 |

Fig. 1 Word sketch for English *catch*, verb (from corpus enTenTen12)

2 The Sketch Engine software: core functions

2.1 The word sketch

The function that gives the Sketch Engine its name is the word sketch: a one-page summary of a word's grammatical and collocational behaviour (Fig. 1).³

This is a feast of information on the word. For *catch* (verb) just looking at the first column (objects of the verb) we immediately see a number of meanings, idioms and set phrases. *We catch a glimpse of* or *catch sight of* something. Fisherman, fishers⁴ and anglers (column 2) *catch fish, trout* and *bass*. You often want to *catch someone's attention*. You sometimes *catch your breath* and things sometimes *catch your eye*. Sportsmen and women, in a range of sports, *catch passes* and *balls*. Things *catch fire*. We all sometimes *catch buses*.

The 'object' column is noise-free, and all items on it are immediately interpretable by a native speaker. The second column, for subject, introduces a couple of complications. *Surprise* relates to the expression *caught by surprise*. *Eye*

³ The examples in this section are all in English, as it is the only language that most readers of the journal share. Later sections will discuss and give examples for a range of Asian languages.

⁴ *Fisher* is a gender-neutral variant of *fisherman* (in addition to its uses in compounds such as *scallop fisher, bottom fisher, the Fisher King*, in the biblical *fisher of men*, and as a common English surname).

| | |
|--------------|---|
| doc#486850 | from the cgprt centre. Why can't dobson get caught with his pants down, or how about falwell |
| doc#1398336 | similar." Looking like a man who has been caught with his pants down and is valiantly trying |
| doc#4315657 | anyhow, and if you're stupid enough to get caught with your digital pants down, woe is you |
| doc#5684214 | without a court order. But I guess he wasn't caught with his pants down in the oval office... |
| doc#9260036 | the fuck is going on?!" he says jealously. Caught with his pants down, Sergio throws Camille |
| doc#11238834 | wright Former Professional not long ago was caught with his pants down and destroyed a lovely |
| doc#11813476 | rescue you and will probably think we've been caught with our pants down. There going to come |
| doc#11986224 | rights because few had been engaged "We were caught with our pants down," he said. "September |
| doc#12513909 | , I'll have to say I was quite literally caught with my pants down. The saga that began |
| doc#12573174 | , or simply knew about it, or was simply caught with its pants down, we may never know. |
| doc#14864966 | Thanks Donna Ah well now,... another one caught with his pants down. But wait !...could |
| doc#16638028 | you routinely make fun of people who get caught with their pants down. I think all of you |
| doc#16662568 | appears that its the Liberals who have been caught with thier pants down in regards to the |

Fig. 2 Concordance for *caught with pants*

and *breath* are objects misanalysed as subjects. *Touchdown catches* is a term from American football: the word sketch succeeds in bringing it to our attention, though *catches* is a noun which has been misanalysed as a verb. *Police* introduces a new meaning of the verb (police catch criminals) and *Anyone* brings to our attention the related pattern *Anyone caught [doing X] will be [punished]*.

The third column, and/or, tells us more about the police and sports meanings. *Overheat* goes with *catch fire*. *Tangle* and *snag* introduce a new meaning where, if a rope or line or piece of cotton or string or wire *catches* with something else, it no longer runs free.

The fourth table brings our attention to the phrasal verbs *catch up*, *catch on*, *catch out*; the fifth, to the reflexive use (*I caught myself wondering...*). The next set of tables show us what we might be caught in (*the crossfire*, *a trap*, *the headlights*), *on* (*videotape*, *CCTV*), *by* and *with* (*your pants down*). The final column takes us back to the police, with people being caught *red-handed* and *unprepared*.

The word sketch can be seen as a draft dictionary entry. The system has worked its way through the corpus to find all the recurring patterns for the word and has organised them, ready for the lexicographer to edit, elucidate, and publish. This is how word sketches have been used since they were first produced.

2.2 Concordance

When looking at a word sketch, a user often wants to find out more: where and how, for example, was *catch* used with *with* and *pant*? They can do this by clicking on the number, and seeing the concordance, as in Fig. 2.

This is usually enough to show why a collocate occurred in a word sketch.

The concordance is the basic tool for anyone working with a corpus. It shows you what is in your corpus. It takes you to the raw data, underlying any analysis. Getting there from a word sketch is just one of the ways of getting to a concordance. The basic method is from the simple search form, as in Fig. 3.

This is modelled on the Google search form. Users like a simple input form, where they put in what they are looking for, and the tool finds it for them. It is for

Fig. 3 Simple search form

Fig. 4 Search hits for simple search *catch fire*

the tool to do its best to understand what it was that the user wanted, and to find it for them. In the case of the Sketch Engine, simple searches are interpreted as follows:

- case-insensitive (so a search for *catch* finds *catch*, *Catch* and *CATCH*)
- as searches for either word form or lemma (so a search for *catch* finds *catch*, *catching*, *catches*, *caught*, and a search for *caught* finds just *caught*), and
- where there is more than one item (with space as separator), a sequence.⁵

These three aspects combine, so a simple search for *catch fire* finds all the hits in Fig. 4.

Users often want more control than the simple search offers. By clicking on ‘Query types’ they see the options as in Fig. 5, and can specify a lemma (with optional word class, e.g. verb, noun, adjective) or a specific phrase or word form (with an option to match for case). ‘Character search’ is designed for languages which do not put spaces between words (Chinese, Japanese, Thai) so users can see a concordance for a character (without having to guess how the text has been segmented into words). CQL is the underlying corpus query language, which technically inclined users can input directly in the CQL box.⁶ Other query types are

⁵ For languages which do not put spaces between words (Chinese, Japanese, Thai), segmentation into words is both a prerequisite for high-quality concordancing, and also a user interface challenge. In the Sketch Engine all corpora for these languages have been pre-processed with language-specific tools to segment the input string into words.

⁶ Corpus query language (CQL) is based on the formalism developed at University of Stuttgart in the 1990s (Christ and Schulze 1994) and widely used in the corpus linguistics community. The Sketch Engine version has been extended (Jakubíček et al. 2010) and is fully documented, with a tutorial, on the website.

Fig. 5 Query types

Fig. 6 Context filters

instead of around it? Isn't he concerned with **catching** his **pants** on fire or getting something
 , go check the guns. I don't want to get **caught** with my **pants** down IF this is true. Not
 a bursting brown paper bag in her arms. **Caught** on his knees, **pants** down, so to speak,
 appropriate at this time to prevent getting " **caught** with our **pants** down". Incidentally, I have
 Telegraph . Put plainly, these people have been **caught** with their **pants** down. They have been caught
 's just an arrogant jackass that has been **caught** with his **pants** down in the gold leasing
 for you to say. You're not the one who was **caught** with their **pants** down." Shit and damnation
 welcome opportunity for the rest of us to **catch** up...and **pant** and clutch our quads in pain

Fig. 7 Concordance for *catch* with context filter *pant* (lemma) within five words *left* or *right*

automatically transformed into CQL queries which are then evaluated by the underlying database engine to obtain the results from the corpus.⁷

Lexicographers often want to home in on a particular pattern of use to explore it further. This can be done with the Context options, as in Fig. 6.

To find all instances of *catch* that have *pant* or *pants* within five words, we search for *catch* with *pant* in as a lemma filter, with results as shown in Fig. 7.

⁷ Most databases are based on relational modelling and queried using Structured Query Language (SQL). However, for text, the sequence of words is the central fact, not the relations, so it is debatable whether SQL databases are suitable. The Sketch Engine is based on its own database management system called Manatee (Rychlý 2000, 2007) devised specifically for corpus linguistics. The web-based front end of Manatee is called Bonito and together with the Corpus Architect module (responsible for building and managing user corpora) these three are the core components of the Sketch Engine.



Fig. 8 Text type options in the Brown Family corpus

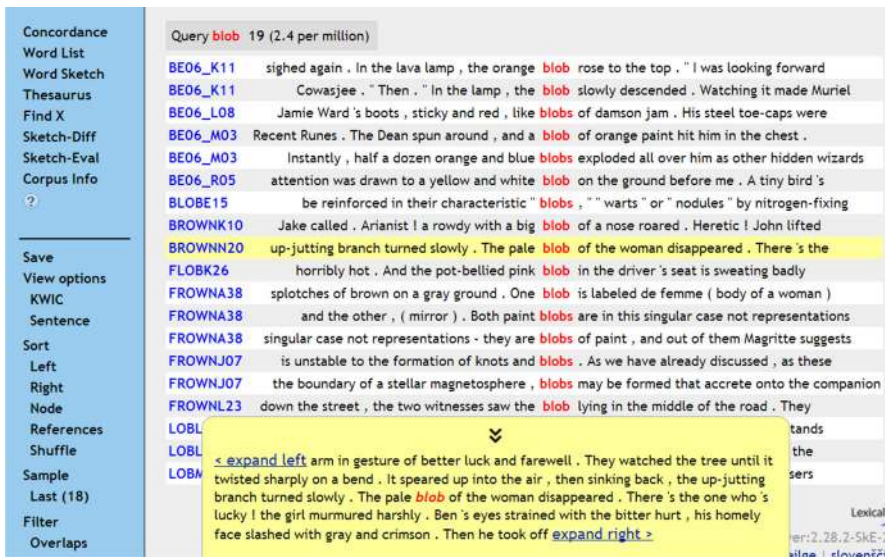


Fig. 9 Concordance display showing how the user can see more context, also showing the ‘Left Hand Menu’ with the range of options for exploring the concordance, and the reference column (in blue) with an identifier for the document that the corpus line came from: clicking on at item in this column will show the metadata for the item

Some corpora have the documents within them classified for text type, for example, the “Brown Family” corpus, comprising the original Brown corpus (American English, 1961) and its various clones (for British or American English and at various date points), all with the same structure and genre distribution. Clicking on Text Type in the concordance form (Fig. 5) shows the form in Fig. 8. The user can limit the search to a particular national variety, time, or genre, by ticking boxes.

Once the user has a concordance, there are many things that can be done with it. It can be sorted, sampled, filtered (for example by Context, or Text Type) or saved. A range of frequency analyses are available, including collocation reports and analysis by text types (where the corpus has text types defined). At the level of the individual hit, the user can click on the search term for more context (see Fig. 9), or on the item in the ‘reference’ column to see the metadata for the item.

tea ^(noun)
enTenTen12 freq = **614,038** (47.3 per million)

| Lemma | Score | Freq |
|------------------|-------|-----------|
| <u>coffee</u> | 0.544 | 783,395 |
| <u>drink</u> | 0.431 | 674,357 |
| <u>juice</u> | 0.416 | 389,262 |
| <u>chocolate</u> | 0.405 | 409,061 |
| <u>wine</u> | 0.39 | 1,048,054 |
| <u>fruit</u> | 0.378 | 858,835 |
| <u>beer</u> | 0.37 | 465,565 |
| <u>beverage</u> | 0.369 | 191,818 |
| <u>milk</u> | 0.368 | 516,439 |
| <u>cream</u> | 0.354 | 543,585 |
| <u>herb</u> | 0.344 | 306,003 |
| <u>honey</u> | 0.328 | 158,877 |
| <u>bean</u> | 0.327 | 263,509 |
| <u>vegetable</u> | 0.325 | 516,918 |
| <u>soup</u> | 0.322 | 188,722 |



Fig. 10 Thesaurus entry for *tea*. In the word cloud, the larger a word, the more similar it is to *tea*

2.3 Thesaurus

The Sketch Engine prepares a ‘distributional thesaurus’ for a corpus. This is a thesaurus created on the basis of common collocation. If two words have many collocates in common, they will appear in each other’s thesaurus entry. It works as follows: if we find instances of both *drink tea* and *drink coffee*, that is one small piece of evidence that *tea* and *coffee* are similar. We can say that they ‘share’ the collocate *drink* (verb), in the OBJECT-OF relation. In a very large computation, for all pairs of words, we compute how many collocates they share, and the ones that share most (after normalisation) are the ones that appear in a word’s thesaurus entry. Distributional thesauruses are a topic of great interest in computational linguistics, and show promise for addressing a range of challenges.

The thesaurus entry for *tea* (in both list and word-cloud form) is shown in Fig. 10.

3 Users and uses

3.1 Lexicography

The first Sketch Engine users were lexicographers, with Macmillan as the first user for the word sketches,⁸ and Oxford University Press as the first for the Sketch Engine.

Lexicography, particularly for English and particularly in the 1980s and 90s, was the driving force in the development of corpus methods and corpus use.

⁸ This was in 1998, for the preparation of the first edition of the Macmillan English Dictionary (Rundell 2002) in a process described in Kilgarriff and Rundell (2002). Thus word sketches are older than the Sketch Engine. The first versions of word sketches were standalone HTML files, one for each word. The integration with a full-function corpus query tool, Manatee/Bonito, to give The Sketch Engine, came later.

Lexicography required very large corpora, so there was evidence even for rare words and phrases. At the time—pre-web or in the web’s infancy, pre “big data”—few others in linguistics or the language professions saw any great need for corpora. The English learners’ dictionaries had a vast and growing market, and were highly profitable, and were competing intensively with each other to produce ‘the best’ dictionary. This was fertile ground for innovation.

Lexicography has continued to be a core use for the Sketch Engine, with four of the five main dictionary publishers in the UK (Cambridge University Press, Harper Collins, Macmillan, Oxford University Press) using it intensively. At CUP and Macmillan, this is just for English; at Collins also for the main European languages, and at OUP also for large bilingual-dictionary projects for Arabic, Chinese and Portuguese.

In the UK, dictionary publishing is dominated by companies (and the commercial wings of University Presses); this is possible largely because there is a very large market. In many countries and for many languages, the curation of the national language is seen as a national project, and most lexicography takes place in academies and national institutes. They form a second group of users for the Sketch Engine. The Sketch Engine is in use at national institutes for Bulgarian, Czech, Dutch,⁹ Estonian, Irish,¹⁰ and Slovak.

3.2 Universities

The Sketch Engine has come out of the academic research world, and, naturally, many of its users are in universities. Within universities, the main kinds of use are

- in linguistics and languages departments: teaching and research
- in computing departments: teaching and research in relation to language technology (also called Natural Language Processing, Computational Linguistics). This is the home area of all Sketch Engine team members
- teaching translation
- discourse analysis: analyses of a particular kind of language for what it tells us about the attitudes, power relations and perspectives of the participants. This kind of work takes place in a range of departments in the humanities and social sciences. Recent examples include the analysis of British newspaper discourse on migrants and migration; portrayal of science in the news; knowledge dissemination through personal blogs.

⁹ Dutch (also called Flemish) is an official language in both the Netherlands and Belgium, and the institute here (INL) is a joint one from both countries.

¹⁰ Much of the development work for the Sketch Engine was undertaken under a contract from Foras na Gaeilge (the official body for the Irish language) in preparation for the creation of a new English-Irish dictionary (<http://www.focloir.ie>). Irish is spoken in both the Irish Republic and Northern Ireland (which is part of the UK) and Foras na Gaeilge is a joint institute of both countries.

3.3 Language teaching

The Sketch Engine is widely used for English Language Teaching and occasionally also for the teaching of other languages including Chinese, Japanese and Arabic.¹¹ The ‘Teaching and Language Corpora’ community has been exploring ways of bringing corpus methods into language-teaching practice since Tim Johns’ work in the 1980s. Johns worked in Birmingham, UK, alongside the COBUILD project for using corpora for lexicography, and the uses of corpora for ELT can be seen as having two parts: indirect use, in the preparation of dictionaries (and coursebooks), as covered above, and direct: in the classroom.

A first ELT coursebook based on the Sketch Engine has recently been published (Thomas 2014).

Countries where the Sketch Engine is widely used in ELT include China, the Czech Republic, Germany, Italy, Japan, Spain and Taiwan as well as the UK.

3.4 Translators

Translators find corpora (of specific domains) useful for identifying the terminology and phraseology of the domain, in the language they are translating into. (They will usually be a native speaker of that language, but will often not know the terms and turns of phrase for specialised areas in which they have a translation task). A number of professional translators are Sketch Engine users.

3.5 Terminologists

In the context of large organisations needing to prepare many documents in multiple languages, consistency is a challenge: in particular, the consistent use of the same term (within each language) for the same concept. It is good practice to develop and maintain a terminology, in which there is an entry for each of the concepts in a domain, with a specification of the term to be used in each language. One of the challenges for terminologists is finding the concepts and terms. The Sketch Engine can be used for term-finding (Kilgarriff 2013). This functionality has been developed in collaboration with the World Intellectual Property Organisation.

3.6 Language technology companies

A word list (with frequencies) for a language is a central resource for almost any language technology application, from speech recognition to spelling correction to text prediction. The corpora in the Sketch Engine provide the raw material, and the software can produce the word lists (and also many other lists: of n-grams,

¹¹ Much of this work takes place in universities, but much also takes place outside (e. g., in language schools) so we treat it as a separate type of use.

Table 1 All the world languages with over 50 million speakers

| Rank | Language | Primary Country | Total countries | Speakers (millions) | SkE status |
|------|---------------------------|--------------------|-----------------|---------------------|------------|
| 1 | Chinese [zho] | China | 33 | 1,197 | Good |
| 2 | Spanish [spa] | Spain | 31 | 406 | Good |
| 3 | English [eng] | United Kingdom | 101 | 335 | Good |
| 4 | Hindi [hin] | India | 4 | 260 | Good |
| 5 | Arabic [ara] | Saudi Arabia | 59 | 223 | Good |
| 6 | Portuguese [por] | Portugal | 11 | 202 | Good |
| 7 | Bengali [ben] | Bangladesh | 4 | 193 | Basic |
| 8 | Russian [rus] | Russian Federation | 16 | 162 | Good |
| 9 | Japanese [jpn] | Japan | 3 | 122 | Good |
| 10 | Javanese [jav] | Indonesia | 3 | 84.3 | No |
| 11 | German [deu] | Germany | 18 | 83.8 | Good |
| 12 | Lahnda [lah] ^a | Pakistan | 7 | 82.7 | No |
| 13 | Telugu [tel] | India | 2 | 74 | Basic |
| 14 | Marathi [mar] | India | 1 | 71.8 | No |
| 15 | Tamil [tam] | India | 6 | 68.8 | Basic |
| 16 | French [fra] | France | 51 | 68.5 | Good |
| 17 | Vietnamese [vie] | Viet Nam | 3 | 67.8 | Good |
| 18 | Korean [kor] | South Korea | 6 | 66.4 | Good |
| 19 | Urdu [urd] | Pakistan | 6 | 63.4 | No |
| 20 | Italian [ita] | Italy | 10 | 61.1 | Good |
| 21 | Malay [msa] | Malaysia | 13 | 59.4 | Basic |
| 22 | Persian [fas] | Iran | 29 | 56.6 | Good |
| 23 | Turkish [tur] | Turkey | 8 | 50.7 | Good |
| 24 | Oriya [ori] | India | 3 | 50.1 | No |

From <http://www.ethnologue.com> (19 April 2014). *Sketch Engine* column added by authors

^a While this is the name for the language as a whole, the better known name for the main dialect is Punjabi

keywords, lemmas, terms) for many languages. Several technology companies have been users of this kind.

4 Languages

The Sketch Engine aims to cover all the large languages of the world, as well as any languages which particular users are asking for.

By a ‘large language’ we mean a language with a large number of speakers. The ethnologue website provides a list of languages sorted by numbers of speakers, as shown in Table 1.

The Sketch Engine has high-level resources for fifteen of these languages (as well as for many smaller ones) and basic resources for a further four. The languages not covered are Javanese (where there is a complex relationship to Bahasa Indonesia, a variety of Malay, for which there is a basic resource) and four of the languages of India and Pakistan (Lahnda/Punjabi, Marathi, Oriya and Urdu).¹²

The prerequisite for a basic resource for a language, is simply, a corpus (plus segmentation tool where there are no spaces between words). A corpus can be collected from the web, using the Corpus Factory (Kilgarriff et al. 2010) or TenTen (Jakubíček et al. 2013) method.

For a high-level resource, further prerequisites are

- a tokeniser (for Chinese and Japanese, usually called a segmenter) to identify the words. In simple cases this might just use spaces between words but many languages have clitics and similar needing language-specific treatment. English is a very simple language in this regard but even there, the hyphen and apostrophe characters, and mixtures of letters and non-letters, present challenges.
- a lemmatiser¹³
- a part-of-speech tagger
- a parser or ‘sketch grammar’.

What is also required is a collaborator. This is a person who speaks the language, and is ideally a computational linguist, and who cares about the quality of the output. They might care because they want to use the corpus in their own (or their group’s) work, or because they developed some of the tools and this is an opportunity to thoroughly test them and to show them (via data processed by them) to the world. The collaboration is crucial: without input from people who speak the language, the Sketch Engine team does not know if what it has done for a language is good. A collaborator is needed to point out mistakes and problems, which can then be addressed.

In the following sections we provide details about the status of Sketch Engine integration of various Asian languages.

4.1 Chinese

The collaboration for Chinese began with Prof Huang Chu-Ren inviting the first author to Taiwan in 2004. (At the time, Huang was Deputy Director of the Linguistics Department at Academia Sinica, Taiwan.) Following that visit, and commercial interest in Chinese in the Sketch Engine from CJKI,¹⁴ the Chinese Gigaword corpus was acquired from the Linguistic Data Consortium, segmented

¹² For Urdu, the relation between the language spoken in India (as a mother tongue) and in Pakistan (as an official language but not the mother tongue of many people) is a particular puzzle. For both Urdu and Punjabi, multiple writing systems are a challenge. For all Indian languages, until quite recently, many Indians would use the web in English: all educated Indians speak English, web searching worked better in English, and not so much could be found in many of the Indian languages. This is changing fast.

¹³ For all of the languages that have been considered except Chinese, where the concept does not apply.

¹⁴ <http://cjki.org>.

攻击 Chinese GigaWord 2 Corpus: Mainland, simplified freq = [6196](#) (24.8 per million)

| Object | 2893 | 3.2 | Subject | 2493 | 3.3 | Modifier | 738 | 2.8 | Modifies | 359 | 0.2 | | | |
|--------|--------------------|------|---------|---------------------|------|----------|--------------------|-------|----------|--------------------|----------------------|--------------------|------------|------------|
| 癌细胞 | 29 | 7.57 | 黑客 | 66 | 9.28 | 无端 | 35 | 10.03 | 对象 | 18 | 6.14 | | | |
| 舰 | 19 | 7.33 | 自杀性 | 20 | 7.33 | 大肆 | 26 | 8.71 | 可能性 | 9 | 4.38 | | | |
| 点多 | 13 | 7.09 | 恶意 | 13 | 7.13 | 轮番 | 12 | 8.27 | 目标 | 58 | 4.15 | | | |
| 潜艇 | 24 | 6.82 | 阿帕奇 | 11 | 7.1 | 蓄意 | 13 | 8.24 | 能力 | 14 | 1.95 | | | |
| 阵型 | 11 | 6.77 | 白棋 | 12 | 6.75 | 肆意 | 13 | 8.14 | 系统 | 12 | 1.42 | | | |
| 直升机 | 36 | 6.66 | 武力 | 35 | 6.74 | 恶意 | 9 | 7.84 | | | | | | |
| 战斗机 | 27 | 6.38 | 导弹 | 110 | 6.72 | 互相 | 26 | 6.63 | | | | | | |
| 对方 | 39 | 6.24 | 进行性 | 8 | 6.64 | 相互 | 58 | 6.27 | | | | | | |
| 弱点 | 11 | 6.2 | 路由器 | 8 | 6.59 | 频频 | 8 | 6.19 | | | | | | |
| 对象 | 22 | 6.04 | 鱼雷 | 8 | 6.57 | 对 | 25 | 6.07 | | | | | | |
| 言论 | 15 | 5.93 | 炮火 | 8 | 6.17 | 加以 | 10 | 5.38 | | | | | | |
| 阵地 | 17 | 5.73 | 黑棋 | 8 | 6.07 | 再次 | 18 | 4.3 | | | | | | |
| | | | | | | | | | | | SentObject of | | 134 | 3.5 |
| | | | | | | | | | | | 用来 | 18 | 6.41 | |
| | | | | | | | | | | | 停止 | 26 | 4.77 | |
| | | | | | | | | | | | 专门 | 10 | 3.01 | |
| | | | | | | | | | | | 继续 | 14 | 1.38 | |
| | | | | | | | | | | | 开始 | 11 | 1.21 | |

Fig. 11 Word sketch for Chinese 攻击 (attack)

and part-of-speech-tagged at Academia Sinica using the tools developed there, and installed into the Sketch Engine. A sketch grammar was developed and word sketches were made available (Huang et al. 2005; Kilgarriff et al. 2005), as illustrated in Fig. 11. They have supported an extensive research programme since (e.g. Chung and Huang 2010; Huang et al. 2014 forthcoming).¹⁵

4.2 Arabic

The collaboration for Arabic is more recent, with the Centre for Computational Linguistics at Columbia University, USA (who prepared MADA + TOKAN, the leading tools for tokenisation, lemmatisation and POS-tagging of Arabic) and Arabic experts elsewhere in the USA, in Saudi Arabia and in the UK. Over a number of years we had received many expressions of interest regarding Arabic in the Sketch Engine. But the language presents a number of challenges:

- there is modern standard Arabic (MSA: the language of the press, education, and officialdom, throughout the Arabic world), Classical Arabic, and the dialects. Most Arabic speakers speak largely their own dialect and are only occasional users of MSA. It is far from obvious what should be included in a corpus;

¹⁵ The resources for Chinese have since been updated, and now the leading Chinese corpus in the Sketch Engine is zhenTen, from the web, and is processed by Stanford tools. The collaboration with Prof. Huang continues.

أخضر arTenTen12 [sample 115M] freq = 9832

| and/or | 472 | 0.7 | adjective-of | 4865 | 4.6 |
|---------|-----|------|--------------|------|-------|
| يايس | 174 | 12.6 | ضوء | 456 | 10.48 |
| أصفر | 52 | 9.52 | شاي | 178 | 10.14 |
| برتقالي | 9 | 8.45 | لون | 300 | 9.42 |
| أزرق | 13 | 8.31 | خط | 355 | 8.94 |
| أحمر | 77 | 7.87 | مسطح | 61 | 8.54 |
| أبيض | 42 | 6.99 | منطقة | 637 | 8.41 |
| أسود | 12 | 6.36 | جبل | 119 | 8.35 |
| | | | حزام | 57 | 8.23 |
| | | | عشب | 51 | 8.1 |
| | | | مسيرة | 83 | 7.76 |
| | | | مساح | 67 | 7.65 |
| | | | نبات | 38 | 7.25 |

Fig. 12 Word sketch for Arabic أخضر (*green*)

- Arabic has many clitics, making tokenisation challenging;
- Arabic is usually written without vowels;
- Arabic has a complex morphological system, with a large share of the vocabulary being the result of derivations according to semi-productive processes. A central issue in Arabic lexicography is whether entry should be based on stems (the traditional approach, giving a smaller number of longer entries) or lemmas (which are closer to dictionary headwords in an English or French dictionary).

It has taken several years to assemble all the pieces required for high-quality resources for Arabic (Arts et al. 2014 forthcoming). An Arabic word sketch is shown in Fig. 12.

4.3 Other Asian languages

For Turkish there were open-source tools available, including a parser, so a Turkish web corpus was processed with that, and the dependency relations which were the output of the parser were used directly to form word sketches (Ambati et al. 2012), see Fig. 13.

For word sketches for Japanese (Srdanovic et al. 2008), Vietnamese (Hà et al. 2012), and Hindi, see Figs. 14, 15 and 16.

For Persian, a very large corpus which had been prepared and parsed at Carnegie Mellon University, USA, was loaded into the Sketch Engine.¹⁶

¹⁶ <http://www.sketchengine.co.uk/documentation/wiki/Corpora/TalkBankPersian>

yürek () TurkishWaC freq = 3483 (82.5 per million)

| MODIF | 906 | 0.9 | OBJ OF | 905 | 1.8 | SUBJ OF | 552 | 1.5 | CLASSIF | 309 | 1.3 |
|-------|-----|------|---------|-----|------|-------------|-----|------|-----------|-----|-------|
| cesur | 24 | 8.52 | dağla | 25 | 9.34 | sızla | 10 | 7.97 | asi | 21 | 10.51 |
| temiz | 17 | 6.18 | burk | 18 | 8.85 | çarp | 13 | 4.65 | yufka | 14 | 9.73 |
| açık | 51 | 5.61 | hopla | 11 | 8.36 | yet | 8 | 4.01 | un | 11 | 3.38 |
| dolu | 12 | 5.07 | sızla | 13 | 8.0 | dayan | 11 | 3.83 | insan | 24 | 2.51 |
| tek | 31 | 4.96 | ferahla | 9 | 7.98 | dur | 17 | 3.0 | in | 8 | 1.99 |
| bütün | 21 | 3.85 | katı | 10 | 7.26 | yan | 19 | 2.73 | in | 11 | 1.53 |
| sev | 9 | 2.85 | parçala | 16 | 6.15 | yaz | 18 | 2.61 | | | |
| yan | 10 | 1.8 | titre | 8 | 6.14 | at | 10 | 1.85 | | | |
| bir | 84 | 1.65 | del | 8 | 6.01 | | | | ABL AD OF | 292 | 9.7 |
| kadar | 14 | 1.35 | yak | 25 | 5.85 | MODIFIER OF | 514 | 0.5 | kutla | 13 | 5.94 |
| gibi | 16 | 1.16 | sar | 11 | 4.88 | insane | 19 | 2.17 | ağla | 10 | 5.73 |
| | | | dol | 12 | 4.4 | | | | inan | 40 | 5.02 |

Fig. 13 Word sketch for Turkish *yürek* (heart)

溜まる JpWaC freq = 2899 (7.1 per million)

| nounが | 1315 | 8.7 | bound_V | 1426 | 5.0 | modifier_Adv | 192 | 4.4 | nounに | 671 | 3.4 |
|-----------|------|------|---------|------|------|--------------|-----|------|-------|-----|------|
| ストレス | 276 | 9.66 | いる | 735 | 1.98 | どんだん | 23 | 6.27 | 中 | 60 | 1.53 |
| 水 | 106 | 6.02 | てる | 194 | 3.36 | かなり | 12 | 3.86 | 底 | 24 | 5.97 |
| 疲れ | 105 | 8.56 | しまう | 117 | 2.71 | いろいろ | 11 | 2.38 | 内 | 17 | 2.29 |
| 疲労 | 39 | 8.18 | くる | 109 | 2.01 | いっぱい | 10 | 4.36 | 間 | 17 | 1.83 |
| フラストレーション | 37 | 9.43 | いく | 104 | 2.26 | また | 9 | 4.01 | うち | 16 | 2.62 |
| 仕事 | 34 | 2.66 | やすい | 38 | 3.58 | 少し | 8 | 3.58 | 体内 | 15 | 7.04 |
| 物 | 27 | 2.8 | すぎる | 15 | 2.62 | だいふ | 5 | 6.18 | 体 | 12 | 2.35 |
| 不満 | 22 | 5.88 | ちゃう | 15 | 2.37 | 相当 | 5 | 3.7 | 上 | 12 | 0.59 |
| ポイント | 18 | 4.54 | 始める | 11 | 1.33 | どりあえず | 4 | 5.62 | 下 | 11 | 1.94 |
| 涙 | 15 | 4.98 | く | 8 | 1.47 | ある程度 | 4 | 5.41 | 肺 | 10 | 7.17 |
| メール | 15 | 3.25 | ゆく | 7 | 2.8 | あまり | 4 | 2.81 | 部 | 10 | 1.63 |
| 空気 | 13 | 4.4 | 過ぎる | 7 | 1.94 | | | | そこ | 9 | 1.41 |
| ガス | 12 | 5.09 | だす | 6 | 2.67 | | | | 部分 | 8 | 1.34 |
| 埃 | 11 | 7.19 | | | | | | | 奥 | 7 | 4.07 |
| 膿 | 10 | 7.56 | | | | | | | 周り | 7 | 3.26 |
| 血 | 10 | 4.27 | | | | | | | 身体 | 7 | 3.06 |

Fig. 14 Word sketches for Japanese 溜まる (accumulate)

tay ⁽⁻ⁿ⁾ VietnameseWaC freq = **63663** (490.5 per million)

| objectArgument | 24,991 | 4.0 | simple_sentence_1 | 18,915 | 3.0 | modifies_N | 2,188 | 3.1 |
|----------------|--------|------|-------------------|--------|------|------------|-------|-------|
| giơ | 986 | 10.0 | ôm | 342 | 8.44 | còng | 262 | 11.3 |
| chấp | 757 | 9.86 | chào | 294 | 8.39 | rãnh | 197 | 11.06 |
| nằm | 1,333 | 9.61 | ra hiệu | 156 | 8.01 | đơ | 79 | 9.85 |
| căm | 906 | 9.34 | nằm | 382 | 7.98 | cút | 57 | 9.35 |
| vẫy | 516 | 9.29 | xách | 123 | 7.51 | chặt | 283 | 8.89 |
| đưa | 2,043 | 8.87 | che | 158 | 7.48 | nhanh | 290 | 8.39 |
| thò | 370 | 8.84 | vuốt | 111 | 7.47 | ngắn | 70 | 8.09 |
| buông | 377 | 8.55 | vẫy | 110 | 7.43 | ghê | 23 | 7.64 |
| vung | 258 | 8.31 | ra | 1,045 | 7.4 | rộng | 77 | 7.53 |
| rửa | 313 | 8.3 | lên | 695 | 7.38 | bồng | 15 | 7.48 |
| đeo | 296 | 8.22 | săn | 111 | 7.28 | le | 14 | 7.36 |
| chia | 233 | 8.2 | xoa | 84 | 7.08 | bao | 19 | 7.34 |

Fig. 15 Word sketches for Vietnamese *tay* (hand)

दिल ^(noun) HindiWaC freq = **16563** (251.8 per million)

| object_of | 2146 | 1.8 | subject_of | 5192 | 2.7 | modifier | 902 | 0.8 | and/or | 550 | 1.2 |
|-----------|------|------|------------|------|------|----------|-----|-------|--------|-----|-------|
| दुखा | 66 | 9.88 | धड़क | 27 | 7.37 | तहे | 107 | 11.76 | दिमाग | 154 | 10.63 |
| बहला | 54 | 9.45 | पड़ | 57 | 7.23 | खुला | 82 | 9.68 | आत्मा | 10 | 6.18 |
| जीत | 244 | 9.35 | बस | 25 | 6.78 | सच्चा | 84 | 9.53 | बच्चा | 22 | 5.17 |
| दहला | 46 | 9.35 | पड़ | 21 | 6.54 | कमजोर | 18 | 7.8 | बात | 11 | 2.21 |
| धड़क | 28 | 8.65 | उछल | 14 | 6.25 | सीधे | 25 | 7.63 | | | |
| दहल | 27 | 8.57 | बसा | 15 | 6.17 | साफ | 16 | 7.31 | | | |
| मचल | 28 | 8.5 | टूट | 24 | 6.1 | ठंडा | 10 | 7.28 | | | |
| खोल | 173 | 8.2 | निकल | 57 | 6.08 | युवा | 22 | 6.85 | | | |
| टूट | 72 | 8.14 | सुन | 49 | 6.05 | खूब | 12 | 6.62 | | | |
| रो | 25 | 7.51 | उतर | 22 | 5.99 | छोटा | 10 | 5.18 | | | |
| उछल | 16 | 7.46 | भर | 68 | 5.95 | पूरा | 22 | 4.87 | | | |
| धड़क | 11 | 7.36 | उठ | 26 | 5.88 | | | | | | |

Fig. 16 Word sketches for Hindi *दिल* (heart)

There are also resources for Azerbaijani, Bengali, Hebrew, Indonesian, Kazakh, Korean, Kyrgyz, Malay, Malayalam, Tajik, Tamil, Telugu, Thai, Tibetan (Garrett et al. 2014 forthcoming), Turkmen, and Uzbek (Baisa and Suchomel 2012).

5 Corpora

“Corpora for all” is the Sketch Engine company tagline: here we give a brief survey of the range of corpora in the Sketch Engine.

Corpora in the Sketch Engine are either owned and managed by the Sketch Engine Team (‘preloaded’ corpora), or are user corpora, owned and managed by the user.

5.1 Preloaded corpora

5.1.1 *General language*

The primary goal is to provide, for each language, a large, recent, general language corpus for the language, processed with high-quality tools for the language, with word sketches, and checked extensively by one or more collaborators. These corpora are for lexicography and general language research, for example into the syntax or morphology of the language. ‘Large’ means at least 50 million words, and for recent work with large languages, several billion. In most cases these are web corpora, as the web is the only place to get material in vast quantity and covering a wide range of text types and domains. In some cases, for example Estonian or Irish, where there is collaboration with an organisation which has gathered a large corpus using some other method, we have combined web-sourced and other material.

These corpora can be kept up-to-date by crawling again and adding new material.

There are large, general-language corpora for sixty languages.

5.1.2 *Parallel*

One central language task is translation. For that, a key resource is the parallel corpus, comprising sets of texts which are translations of each other (or, are both translations of the same source). Parallel concordancing, as in Fig. 17, is where a user inputs a search term in one language and sees pairs of sentence: those with the matching term in the first language, and the corresponding sentence in the target language.

In the Sketch Engine there are data for 300 language pairs. These data are from two main sources: EUROPARL and OPUS. EUROPARL comprises speeches made at the European parliament, which have been translated into 21 official European Union languages (Koehn 2005).¹⁷ The OPUS data, a collection of parallel corpora collected in the OPUS project, are made available on its website.¹⁸ It comprises

¹⁷ The original release was for fewer languages. For recent releases see <http://www.statmt.org/europarl/>.

¹⁸ <http://opus.lingfil.uu.se/>.

| OPUS2 Chinese Simplified | OPUS2 English |
|--|---|
| 我想我能叫 Lee Marvin 還有 Van Cleef 的 微笑 | I guess I could be Lee Marvin with a Van Cleef smile . |
| 好了，忘了吧，讓我們 微笑 的 ... | " Let' s just sweep everything under the rug with a pat on the back, a smile and don' t forget the glass of champagne to take the edge off. " |
| 我希望能 在 普通 民 众 的 脸 上 看 到 微笑 | I hope the average Joe will be able to keep the smile on his face. |
| 我 相 信 她 一 定 在 上 面 向 我 们 微笑 | Lily: I' m sure she' s smiling down on us right now. |
| 如 果 我 微笑 着 | What if I were smiling and running into your arms? |
| 她 躺 在 我 胸 口 微笑 的 樣 子 | My Maria and ... her little smile when she laid on my chest. |
| “ 向 上 微笑 ” | And why is no one ever says, " I think he' s down there now, smiling up at us. " |
| ?? 错，我 还 是 值得 别人 对 我 微笑 并且 对 我 说 我 很 美?? 动 人 的 | Yeah, I deserve to have somebody smile at me and tell me I' m pretty. |
| 愿 神 对 你 微笑，再 见 了 | Well, may the gods smile on you. See you around. |
| 并且 永 记 住 微笑 比 子 弹 便宜 | And always remember that a smile is cheaper than a bullet. |

Fig. 17 Parallel concordance, Chinese and English, 微笑 and *smile*

many different parts, two of the largest (for most language pairs) being documents from the United Nations, and Open Subtitles.¹⁹ Figure 17 shows subtitle data.

In addition to all of the functionality shown so far, an extra option for parallel data is to search simultaneously in both languages: Fig. 17 shows the output when 微笑 is searched on the Chinese side, *smile* on the English.

5.1.3 Second/foreign language learning and teaching

In the context of language learning, two central questions arise:

- what are learners saying and writing?
- what should they be saying and writing?

For the first, there are learner corpora.²⁰ Learner corpora are valuable for finding out what learners, at various levels, do, and for research into the process of language learning as well as the practicalities of curricula, course development, and testing. In the Sketch Engine there are learner corpora for Slovene, Czech and English.²¹

For the second, the general answer is “the language”, and general language corpora meet that need. But there is also a more specific answer: one large population of language learners are learning English, and would like to study at an English-medium university. Thus their target is the English that is spoken in

¹⁹ <http://www.opensubtitles.org/>.

²⁰ Sketch Engine features designed for learner corpora are described in Kosem et al. (2013). For general information on learner corpora see <http://www.lernercorpusassociation.org/>.

²¹ The Cambridge Learner Corpus, the largest learner corpus for English, is in the Sketch Engine and is used extensively by Cambridge University Press and researchers and textbook authors who publish with them. However it is not accessible to SkE users without a CUP affiliation.

seminars and written in University-level essays, by accomplished English speakers. The British Academic Spoken English (BASE) and British Academic Written English (BAWE) corpora have been created as samples of these target varieties.²²

5.1.4 *Historical*

A central topic for linguists is language development and change. Corpora looking back over the history of a language, and supporting this kind of research, include LatinISE (of Latin from the third century B. C. to the twentieth century A.D., McGillivray and Kilgarriff 2013), GermanC (of German from the seventeenth and eighteenth centuries; Scheible et al. 2011) and English Dialogues Corpus (sixteenth–eighteenth centuries; Culpeper and Kytö 2010).

For the Arabic world and Islam, the language has a special role. It is the language of the Quran and of the culture that the region shares. The different countries each have their own dialect, and the lingua franca, MSA, is closer to classical Arabic than to the dialects. The King Saud University Corpus of Classical Arabic (KSUCCA²³) brings together many of the central texts of this language, culture and religion, including the Quran and the Hadith.

5.1.5 *Learning to speak*

Since 1984, the CHILDES and Talkbank projects, based at Carnegie Mellon University, have been gathering child–adult conversations.²⁴ They are largely between babies and young children and their carers (with many of the carers being linguists, who have taken on the recording and transcription of the data). All are available as transcripts, and many also as audio or video. The data can be explored on the Talkbank website as well as the Sketch Engine: the two websites are complementary, with Talkbank expecting the user to be a developmental or general linguist, and the Sketch Engine expecting them to have a corpus orientation. There is a CHILDES corpus in the Sketch Engine for 22 languages, varying in size from a few thousand words to, for English, 23 million.

5.1.6 *Learning to read and write*

Educators, children’s authors and publishers, and linguists and psychologists studying the process of learning to read, are interested in the language that schoolchildren read and write. So are producers of children’s dictionaries. The Education division of Oxford University Press has created the Oxford Children’s Corpus (Wild et al. 2013), comprising both material written for children (largely stories, many being titles published by OUP) and stories written by children. This second part resulted from a competition led by the top UK disc jockey Chris Evans,

²² <http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/>,
<http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/>.

²³ <http://mahaalrabiah.wordpress.com/category/ksucca/>.

²⁴ <http://childes.psy.cmu.edu/>, <http://www.talkbank.org>.

who, from his show on BBC Radio 2, invited children to write a 500-word story and send it into him. In 2014, 115,000 British children did so. The BBC then made the data available to OUP for linguistic research.²⁵

The size of the corpus, as at April 2014, is 115 million words.

5.1.7 Reference corpora

The Brown corpus was central to the development of corpus linguistics. It was one million words, comprising five hundred 2,000-word samples from 13 different genres, all of American English published in 1961. It has played a huge role as a point of reference ever since and has spawned ‘Brown family’ corpora for British and American English, at a number of time points (see Fig. 5).

Another key reference corpus for English is the British National Corpus (Burnard 1995).

5.1.8 Sociolinguistics

Sociolinguists are interested in how language varies between social groups, across age groups, with movements of populations, and between communities. A corpus designed to study these topics is the London English corpus (Kerswill et al. 2013).

5.2 User corpora

As well as preloaded corpora (managed by the Sketch Engine team) users can upload, build, process, share and explore their own corpora.

Where a user already has a corpus, they can upload it and install it, via a simple web interface. The source documents can be in any of the common formats (doc, html, pdf, txt, tnx) and may also be compressed and/or archived (.zip, .gz, .bz2, .tar). All of these formats are then converted to plain text (.txt). If the data are already annotated (perhaps with part-of-speech tags, or lemmas, or for discourse function, etc.) then it needs to be in the Sketch Engine’s input format, ‘vertical’ text, as documented in the Sketch Engine help pages. The user can then manage their own corpora, including adding more data, deleting, and processing (see below) as well as using them for their research via the core Sketch Engine functions (as in Sect. 2).

BootCaT (Baroni and Bernardini 2004) is a procedure for building a corpus, starting from a set of ‘seed words’, by making tuples (typically triples) of the seed words, sending each tuple as a query to a search engine, and then gathering the web pages that the search engine finds. When applied to a specialist domain, with seed words from that specialist domain, it turns out to be a remarkably efficient way of discovering the terminology and phraseology of that domain. The Sketch Engine includes an implementation called WebBootCaT. See Fig. 18 for the WebBootCaT

²⁵ See <http://global.oup.com/uk/pressreleases/500words/>. Access to the corpus is restricted to OUP staff and collaborators. Word lists based on the corpus may be made available; applications to OUP.

WebBootCaT: Create corpus

[Get seed words from Wikipedia](#)

Corpus name

Language English ▾
Creating BootCaT corpora is available only for those language which we can at least tokenise. All such languages are listed here.

Build word sketches
This option only has effect if a pre-loaded sketch grammar is available for the selected language.

Input type
 Seed words
 URLs
Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

Seed words

Random tuples will be selected from the seed words to query a search engine. Input 3 to 20 words or multiword expressions. Use space as separator. Enclose multiword expressions into quotes ("").

Compile corpus when finished
Automatically compile corpus when WebBootCaT processing is finished.

[Show advanced options](#)

Fig. 18 WebBootCaT form

form and Fig. 19 for the keywords and terms found, fully automatically and in a few minutes, for the vulcanology domain.

More can be done with a user corpus if it is accurately tokenised, lemmatised and part-of-speech tagged. Tools for these processes are language-specific. For 11 major world languages, tools have been identified, licenced (where necessary) and installed. Some users have used the Sketch Engine explicitly for this service: they can upload plain text, get it processed, and then download the processed data (in vertical format), perhaps for further annotation and re-uploading.

Processing options are steadily being added for more languages.

There is ‘access control’ for user corpora. By default, the only user who can see a corpus is the person who created it, but they can give access to others (at a web interface). This allows a teacher to give access to a corpus that they have prepared, to their students, or a researcher to share their corpus with their colleagues.

Vulcanology_en: Extracted keywords and terms

| Keywords | | Terms |
|--|--|---|
| <input type="checkbox"/> lava (62.3, 3349) | <input type="checkbox"/> hawaiian (5.9, 270) | <input type="checkbox"/> lava flow (9.0) |
| <input type="checkbox"/> eruption (55.0, 2939) | <input type="checkbox"/> kilauea (5.9, 261) | <input type="checkbox"/> cinder cone (6.2) |
| <input type="checkbox"/> volcano (45.9, 2463) | <input type="checkbox"/> kilometers (5.9, 287) | <input type="checkbox"/> pyroclastic flow (5.7) |
| <input type="checkbox"/> ash (42.5, 2333) | <input type="checkbox"/> active (5.8, 492) | <input type="checkbox"/> volcanic ash (5.4) |
| <input type="checkbox"/> eruptions (41.2, 2167) | <input type="checkbox"/> miles (5.7, 488) | <input type="checkbox"/> eruption column (4.3) |
| <input type="checkbox"/> volcanic (36.7, 1979) | <input type="checkbox"/> basaltic (5.7, 248) | <input type="checkbox"/> magma chamber (4.3) |
| <input type="checkbox"/> magma (36.0, 1870) | <input type="checkbox"/> molten (5.5, 245) | <input type="checkbox"/> explosive eruption (4.1) |
| <input type="checkbox"/> vent (29.3, 1547) | <input type="checkbox"/> mantle (5.5, 251) | <input type="checkbox"/> volcanic eruption (4.0) |
| <input type="checkbox"/> crater (25.5, 1385) | <input type="checkbox"/> particles (5.5, 283) | <input type="checkbox"/> ash plume (4.0) |
| <input type="checkbox"/> flows (24.9, 1449) | <input type="checkbox"/> volume (5.2, 375) | <input type="checkbox"/> differential flow (3.9) |
| <input type="checkbox"/> volcanoes (23.3, 1208) | <input type="checkbox"/> lavas (5.2, 222) | <input type="checkbox"/> lava lake (3.8) |
| <input type="checkbox"/> flow (21.7, 1632) | <input type="checkbox"/> steam (5.1, 252) | <input type="checkbox"/> lava dome (3.1) |
| <input type="checkbox"/> pyroclastic (20.4, 1029) | <input type="checkbox"/> seismic (5.1, 228) | <input type="checkbox"/> volcanic activity (3.1) |
| <input type="checkbox"/> tephra (20.4, 1026) | <input type="checkbox"/> viscosity (5.1, 221) | <input type="checkbox"/> rift zone (3.1) |
| <input type="checkbox"/> caldera (20.2, 1027) | <input type="checkbox"/> occurred (5.1, 344) | <input type="checkbox"/> strombolian activity (3.0) |
| <input type="checkbox"/> cone (19.3, 1005) | <input type="checkbox"/> crust (5.0, 223) | <input type="checkbox"/> east rift (2.9) |
| <input type="checkbox"/> plume (17.8, 906) | <input type="checkbox"/> explosion (4.9, 240) | <input type="checkbox"/> molten rock (2.7) |
| <input type="checkbox"/> explosive (17.6, 949) | <input type="checkbox"/> ejected (4.9, 209) | <input type="checkbox"/> pyroclastic material (2.7) |
| <input type="checkbox"/> vents (12.5, 617) | <input type="checkbox"/> earthquakes (4.8, 214) | <input type="checkbox"/> se crater (2.6) |
| <input type="checkbox"/> cinder (12.3, 601) | <input type="checkbox"/> zone (4.8, 285) | <input type="checkbox"/> explosive activity (2.6) |
| <input type="checkbox"/> strombolian (11.8, 573) | <input type="checkbox"/> tremor (4.8, 204) | <input type="checkbox"/> east rift zone (2.5) |
| <input type="checkbox"/> activity (11.4, 1126) | <input type="checkbox"/> shield (4.8, 229) | <input type="checkbox"/> flow vent (2.5) |

Fig. 19 Terminology in the volcanoes domain, as extracted from BootCaTted Vulcanology corpus. Items in *green* were input seeds. Number (in *blue*) can be clicked to see concordances

6 New functionality

At ten years old, the Sketch Engine is mature software. There has been a steady stream of new functionality as well as bug-fixes and improvements to usability. Recent usability improvements include

- a ‘breadcrumb trail’ to show a user how they got to the concordance they are looking at, which might be the results of an original search plus sorting, filtering and sampling. This was a response to user feedback that it was easy to lose track of what a particular concordance was
- ‘more data’ and ‘less data’ buttons for work sketches. The number of collocates shown in a word sketch is defined by three parameters: a frequency threshold, a salience threshold and a limit to the number of collocates per list. Users

sometimes want to see more collocates than they have been shown in the first instance—and sometimes they feel overwhelmed and want to see less. But if they start considering the three parameters, they are bamboozled. Hence the ‘more data’ and ‘less data’ buttons

- thesaurus word clouds: see Fig. 9.

Other additions to functionality, which complement the core functions described in Sect. 2 and the preloaded and user corpora as described in Sect. 5, include the API, GDEX, bilingual sketches, keywords and ‘comparing corpora’, terminology, and localisation.

6.1 API

A simple JSON API allows other programmes to access word sketches, collocations, thesaurus entries and to find the terminology in a document.

6.2 GDEX

Dictionary users like examples. This is a clear finding of dictionary user research (Frankenberg Garcia 2014). Where the dictionary is to be published on paper, not many examples can be offered owing to space limitations. With electronic dictionaries, that constraint disappears. The constraint becomes, rather the editorial time needed to prepare them. There are already compelling linguistic reasons for taking examples from corpora rather than inventing them (Hanks 2012): could the corpus software not merely find the examples for a word, but automatically find the good ones, for using as dictionary examples?

A Good Dictionary Examples (GDEX) function was added to the Sketch Engine in 2008 and has had many enthusiastic users. It was originally applied to English (Kilgarriff et al. 2008) and has since been used for a number of languages including Slovene (Kosem et al. 2011). It works by sorting a concordance, so the corpus lines judged best by the algorithm are shown first. Then the lexicographer should not have to read many of them before finding a good one. The same core technique has also been used to score documents and to exclude low-scoring documents from a corpus entirely.

The critical outstanding question, for dictionary publishers, is this: can GDEX work well enough, so that example sentences can be added to dictionary entries without an editor needing to check them first? This is a goal, but a number of obstacles stand in the way. First the corpus needs to be very big, to provide plenty of examples for the algorithm to choose amongst, and usually the only way to get a very large corpus is from the web. But web corpora contain web spam, which sometimes makes it past all other filters and makes bad dictionary examples.

Second: parsnips. Parsnips is an acronym for the potentially offensive topics which teaching materials, which will be seen across the globe, by all communities and cultures, might be wise to avoid. It stands for Politics Alcohol Religion Sex Narcotics Isms Pork (as a stand-in for various foods which are taboo in various cultures). A second current challenge is to scrub the data clean of parsnips.

6.3 Bilingual sketches

Where monolingual lexicographers appreciate monolingual word sketches, bilingual lexicographers would like bilingual ones. They have recently been developed (Baisa et al. 2014) and are currently being rolled out for more language pairs.

6.4 Keywords and corpus comparison

Where there are a number of corpora available for a language, the question arises, “how do they compare?” This has been the central research question for the first author for some years (Kilgarriff 2001, 2012) and the Sketch Engine supports a range of comparisons, quantitative and qualitative, between any pair of same-language corpora: see Kilgarriff (2012) for English and Czech, Kilgarriff and Renau (2013) for Spanish.

6.5 Terminology

To find the terminology of a domain, in a language, the requirements are

- a domain corpus
- a reference corpus
- a grammar for terms
- a lemmatiser, part-of-speech tagger, and parser (to find linguistic units with the grammatical shape that makes it possible for them to be terms)
- a statistic (to identify the term candidates that are most distinctive of the domain in contrast to the reference material).

The Sketch Engine has most of these pieces in place. Users can upload their domain corpus, or build one using WebBootCaT. Reference corpora are available for 60 languages. There are already grammars for the word sketches, which can be adapted to provide term grammars. The parsing machinery is in place, and, as discussed above, for a growing number of languages, language-specific processing tools are installed and ready to use. The statistic used to identify keywords is also suitable for identifying terms.

Term-finding functionality is now available in the Sketch Engine, as illustrated in Fig. 19, for ten languages.

6.6 Localisation

Machinery to support the localisation of the interface has been added. Currently, the interface can be seen in Czech, Chinese, English, Irish, Slovene, and Croatian.

7 Related work

The Sketch Engine is both a corpus query tool and a web service; the web service includes corpus building and management. We take each in turn.

7.1 Corpus query tools

Software tools for corpus exploration fall into two categories: those designed for installation on each computer where they are used, and those designed for installation on a server.

Widely used tools for local installation include (in order of their invention), Monoconc/Paraconc (since 1995), WordSmith (since 1996), Antconc (Anthony 2004; since 2004) and Concgram (Greaves 2009; since 2005). Antconc is free, and the other three are commercial products. All have many enthusiastic users. All can be used over a network, but this is not their normal mode.

Amongst tools for use over a network, the IMS Corpus Workbench has pride of place. IMS is the Institut für Maschinelle Sprachverarbeitung at the University of Stuttgart, where the tool was developed in the early 1990s (Christ and Schulze 1994). (It is also often referred to as CWB, Stuttgart tools, or CQP, for its “Corpus Query Processor”.) It has been widely used and has a community of developers working with it. The original version was pre-web, and the envisaged network was within a University. A central question since is as to how it can be made to work well on the web. The usual solution has been that it provides a back end, and then a number of front ends have been prepared. CQPWeb (Hardie 2012), for example, combines the IMS Corpus Workbench back end with a MySQL database.

The Stuttgart tools and CQPWeb are both free and open-source, and the community of developers for corpus software has a strong commitment to open source. While the Sketch Engine is not open source, as this could undermine its viability as a business, a version of it, NoSketchEngine, is open source.²⁶

The functionality of all of these tools (and most of those covered below) comprises a concordancer, plus various ways to manipulate concordances, plus a range of summary reports. There is little disagreement about the value of the various reports, and the functionality differences lie rather in how much time and motivation the developers have had to develop more functions. As one of the more mature tools, working commercially with a support and development team of seven, the Sketch Engine has more functions than most.

7.2 Corpus websites and services

There are a small number of corpus websites for multiple languages and a large number for a single language (and usually, a single corpus). We review those that cover multiple languages in some detail below. We do not cover the single-language ones: there are too many, many of which are short-lived. It is *de rigueur* for any corpus project to make its corpus available over the web, and this is typically done on a dedicated website, sometimes using the Stuttgart tools as the back end and sometimes using software developed as part of the project. Such projects are often national projects, and one advantage is often that the interface is in the same language that the corpus is a corpus of. For scholars of that language who may not

²⁶ NoSketchEngine comprises current versions of Manatee and Bonito, but without word sketches and the functionalities built on them. See <http://nlp.fi.muni.cz/trac/noske>.

be at ease in English, this may be a major advantage. A disadvantage of developing software afresh is that the software will be new: it is likely to be less robust, with less functionality, than mature systems. The funding will end and then it will be hard to maintain. Growing numbers of corpus developers are taking the route of making their corpus available in the Sketch Engine.²⁷

7.2.1 *Corpus websites for multiple languages*

Mark Davies's website at Brigham Young University²⁸ offers corpora for English, Spanish and Portuguese. The resources for English are outstanding, supporting the exploration of the behaviour of words and phrases across time, genre, and regional varieties (Davies 2009). The system is fast and reliable.

Uwe Quasthoff and colleagues in Leipzig have crawled the web for corpora of 229 languages and made them searchable at their Wortschatz website (Quasthoff et al. 2006).²⁹ The website is in German. Within Germany, this is a very widely used site: it serves as a main reference for language questions from laypeople.

Eckhard Bick has focussed on syntax and parsing. The Visual Interactive Syntax Learning website³⁰ has corpora which are often modest in size but are parsed. The website has games and quizzes to support language education, as well as Deepdict, comprising word-sketch-like reports, for nine languages (Bick 2009).

The OPUS project (Tiedemann and Nygaard 2004) has gathered parallel corpora and organised them so they are both searchable on the website (with the Stuttgart tools as the back end) and also downloadable, for use in other research and software (including the Sketch Engine; many of the parallel corpora in the Sketch Engine were taken from the OPUS site).³¹

At the University of Leeds in the UK, Serge Sharoff makes web corpora for 13 languages available to all for searching, again using the Stuttgart tools back end (Sharoff 2006).³²

All of these sites are free to use. This is in contrast to the Sketch Engine. Most are based in Universities and are supported via research grants and academic salaries. While, naturally, most people would rather not pay (other than via their taxes), the commercial model has advantages. There is an income stream to support the maintenance and development of the software and web service for the long term, and customers with particular requirements can get what they need, by paying for it.

Google and other search engines do a similar job to a corpus website: they allow the user to find many instances of a word, in context, as a dataset for further study, and they do it fast. Where the user knows of no corpus for the language, or the item they are searching for is rare so not enough data are available via dedicated corpus

²⁷ This is the history of most of the Sketch Engine interface localisations.

²⁸ <http://corpus.byu.edu>.

²⁹ <http://wortschatz.uni-leipzig.de/>.

³⁰ <http://beta.visl.sdu.dk/>.

³¹ <http://opus.lingfil.uu.se/>.

³² <http://corpus.leeds.ac.uk/internet.html>.

linguistic tools, Google may be the best tool to use. For a discussion of the use of search engines for corpus research see Kilgarriff (2007).

A possibility lying between the search engine and a corpus tool is the metasearch engine, in which a corpus tool takes a user's query, passes it on to Google or another search engine, receives the results, and filters and displays them in ways that are useful for language researchers. The best known tool of this kind is Webcorp (Renouf et al. 2006).³³

Other corpus-like websites, mentioned here for completeness, are:

- Wikipedia : the wikipedia for a language is a convenient corpus for that language, as used, for example, as a starter corpus in Kilgarriff et al. (2010).
- Project Gutenberg³⁴
- Google books³⁵
- the Linguistic Data Consortium³⁶ and European Language Resource Association,³⁷ for catalogues of available resources, including corpora of various kinds for many languages
- Linguee,³⁸ Webitext,³⁹ parallel concordancers offered as a service to translators.

7.3 Tools for corpus building and annotation

The BootCaT procedure is described above. There are a number of implementations, including one from the University of Bologna group where the idea was originally developed.⁴⁰

Several groups have developed pipelines for web corpus building. The steps are

- web crawling
- removing duplicates
- 'cleaning' to remove non-text material
- language identification
- linguistic processing (tokenisation, possibly also lemmatisation, part-of-speech tagging, parsing).

The pipeline used by the Sketch Engine team uses three tools which were developed within the group and have now been published as open-source software: spiderling (Suchomel and Pomikálek 2012) for crawling, onion for deduplication, justext for cleaning (both Pomikálek 2011). Other pipelines, with similar philosophy and components, have been developed in Bologna (Baroni et al. 2009), Leipzig (Biemann et al. 2004) and Berlin (Schäfer and Bildhauer 2013).

³³ <http://www.webcorp.org.uk/live/>.

³⁴ <http://www.gutenberg.org>.

³⁵ <http://books.google.com>.

³⁶ <https://www ldc.upenn.edu/>.

³⁷ <http://www.elra.info/>.

³⁸ <http://www.linguee.com>.

³⁹ <http://webitext.com>.

⁴⁰ <http://bootcat.sslmit.unibo.it/>.

Annotating a corpus with human input (as distinct from a fully automatic process) is supported in a limited way in the Sketch Engine, via facilities developed for Hanks's Corpus Pattern Analysis (Hanks 2008).⁴¹ Many tools have been developed specifically for manual and semi-automatic corpus annotation, leading examples being the UAM tool (O'Donnell 2008) and the Groningen Meaning Bank tool (Basile et al. 2012).⁴²

8 Conclusion

The Sketch Engine is a leading corpus tool (both in the sense of 'corpus query tool' and in the sense of 'corpus web service'). It is now 10 years old: a 10-year period that has seen revolutions in connectivity, devices, and dictionary publishing, and the worldwide spread of corpus methods in dictionary-making. It is mature software offering a wide range of functions, with the web service offering many corpora for many languages, as well as services for corpus building and maintenance.

In this paper we have described word sketches, concordancing, and the thesaurus (Sect. 2), the different kinds of user (Sect. 3), and approaches to working with many different languages (Sect. 4). Section 5 reviewed the kinds of corpora available in the Sketch Engine, including user corpora and the ways of building and working with them. Section 6 gives a brief tour of some of the innovations and new reports offered in the past few years. In Sect. 7 we reviewed related work.

As the strapline has it, 'corpora for all!'

Acknowledgments This work has been partly supported by the Ministry of the Interior of the Czech Republic, in project VF20102014003.

References

- Ambati, B.R., S. Reddy, and A. Kilgarriff. 2012. Word sketches for Turkish. In *Proc LREC*, 2945–2950. Istanbul.
- Anthony, L. 2004. AntConc: a learner and classroom friendly, multi-platform corpus analysis toolkit. In *Proc IWLeL*, 7–13.
- Arts, T., ed. 2014. *Oxford Arabic Dictionary*. Oxford: Oxford University Press.
- Arts, T., Y. Belinkov, N. Habash, A. Kilgarriff, and V. Suchomel. 2014 (forthcoming). arTenTen and word sketches for Arabic. *Journal of King Saud University: Computing and Information Science*. Special issue on Arabic natural language processing.
- Baisa, V., M. Jakubíček, A. Kilgarriff, V. Kovář, and P. Rychlý. 2014. Bilingual word sketches: the *translate* button. In *Proc EURALEX*, Bolzano/Bozen
- Baisa, V., and V. Suchomel. 2012. Large corpora for Turkic languages and unsupervised morphological analysis. In *Proc LREC*, Istanbul
- Basile, V., J. Bos, K. Evang, and N. Venhuizen. 2012. Developing a large semantically annotated corpus. In *LREC* vol. 12, 3196–3200.
- Baroni, M., and S. Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proc LREC*, Lisbon

⁴¹ <http://pdev.org.uk>.

⁴² <http://gmb.let.rug.nl/>.

- Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3): 209–226.
- Bick, E. 2009. DeepDict—a graphical corpus-based dictionary of word relations. In *Proc NODALIDA*, Vol. 4, 268–271.
- Biemann, C., S. Bordag, G. Heyer, U. Quasthoff, and C. Wolff. 2004. Language-independent methods for compiling monolingual lexical data. In *Computational linguistics and intelligent text processing*, 217–228. Berlin Heidelberg: Springer.
- Burnard, L. 1995. The BNC reference manual.
- Christ, O., and M. Schulze. 1994. The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual. University of Stuttgart.
- Chung, S.-F., and C.-R. Huang. 2010. Using collocations to establish the source domains of conceptual metaphors. *Journal of Chinese Linguistics* 38(2): 183–223.
- Culpeper, J., and M. Kytö. 2010. *Early Modern English dialogues: spoken interaction as writing*. Cambridge: Cambridge University Press.
- Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2): 159–190.
- Frankenberg Garcia, A. 2014. The use of corpus examples for language comprehension and production. ReCALL.
- Garrett, E., N.W. Hill, A. Kilgarriff, R. Vadlapudi, and A. Zadoks. 2014. The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries. In *The Third International Conference on Tibetan Language*, eds. Tuttle, Gya, Dare and Wilber, New York: Trace Foundation (forthcoming).
- Greaves, C. 2009. ConcGram 1.0: a phraseological search engine. John Benjamins.
- Hanks, P. 2008. Mapping meaning onto use: a Pattern Dictionary of English Verbs. In *Proc ACL*, Utah.
- Hanks, P. 2012. The corpus revolution in lexicography. *International Journal of Lexicography* 25(4): 398–436.
- Hardie, A. 2012. CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics* 17(3): 380–409.
- Hà, P.T., N.T.M. Huyền, L.H. Phương, and A. Kilgarriff. 2012. Nghiên cứu từ vựng tiếng Việt với hệ thống sketch engine. *Tạp chí Tin học và Điều khiển học* 27(3): 206–218.
- Huang, C.-R., K.-J. Chen, and Q.-X. Lai. 1997. (黃居仁, 陳克健, 賴慶雄。編著)。國語日報量詞典 *Mandarin Daily Dictionary of Chinese Classifiers*. Taipei: Mandarin Daily Press.
- Huang, C.-R., J.-F. Hong, W.-Y. Ma, and P. Šimon. 2014. From corpus to grammar: automatic extraction of grammatical relations from annotated corpus. In *T'sou and Kwong Eds. Linguistic Corpus and Corpus Linguistics in the Chinese Context. Journal of Chinese Linguistics Monograph*. Hong Kong: Chinese University of Hong Kong Press, (forthcoming).
- Huang, C.-R., A. Kilgarriff, Y. Wu, C.M. Chiu, S. Smith, P. Rychly, M.H. Bai, and K.-J. Chen. 2005. Chinese Sketch Engine and the extraction of grammatical collocations. In *Proc Fourth SIGHAN Workshop on Chinese Language Processing*, 48–55.
- Jakubíček, M., A. Kilgarriff, D. McCarthy, and P. Rychlý. 2010. Fast syntactic searching in very large corpora for many languages. In *Proc PACLIC*, Vol. 24, 741–747, Japan.
- Jakubíček, M., A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel. 2013. *The TenTen corpus family*. Lancaster: In Proc. Int. Conf. on Corpus Linguistics.
- Kerswill, P., J. Cheshire, S. Fox, and E. Torgersen. 2013. English as a contact language: the role of children and adolescents. In *English as a Contact Language*, 258.
- Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1): 97–133.
- Kilgarriff, A. 2007. Googleology is bad science. *Computational linguistics* 33(1): 147–151.
- Kilgarriff, A. 2012. Getting to know your corpus. In *Text, Speech and Dialogue*, 3–15. Berlin Heidelberg: Springer.
- Kilgarriff, A. 2013. Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine. In *Proc ASLIB 35th Translating and the Computer Conference*, London.
- Kilgarriff, A., and M. Rundell. 2002. Lexical Profiling Software and its lexicographic applications: a case study. In *Proc EURALEX*. Copenhagen, Denmark.
- Kilgarriff, A., P. Rychlý, S. Smrz, and D. Tugwell. 2004. The Sketch Engine. In *Proc Eleventh EURALEX International Congress*. Lorient, France.

- Kilgarriff, A., C.R. Huang, P. Rychlý, S. Smith, and D. Tugwell. 2005. Chinese word sketches. In *Proc ASIALEX 2005: Words in Asian cultural context*. Singapore.
- Kilgarriff, A., M. Husák, K. McAdam, M. Rundell, and P. Rychlý. 2008. GDEX: automatically finding good dictionary examples in a corpus. In *Proc. Euralex*. Barcelona
- Kilgarriff, A., and I. Renau. 2013. esTenTen, a Vast Web Corpus of Peninsular and American Spanish. *Procedia Social and Behavioral Sciences* 95: 12–19.
- Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. *Proc MT summit* 5: 79–86.
- Kosem, I., M. Husak, and D. McCarthy. 2011. GDEX for Slovene. In *Proceedings of eLex*, 151–159. Bled, Slovenia.
- Kosem, I., V. Baisa, V. Kovář, and A. Kilgarriff. 2013. *User-friendly interface of error/correction-annotated corpus for both teachers and researchers*. Solstrand: Proc Learner Corpus Research.
- McGillivray, B., and A. Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In *New Methods in Historical Corpora*, Bennett, P.D. ed. Vol 3. BoD—books on demand.
- O'Donnell, M. 2008. Demonstration of the UAM CorpusTool for text and image annotation. In *Proc 46th ACL: Demo Session*, 13–16. Association for computational linguistics.
- Pomikálek, J. 2011. Removing boilerplate and duplicate content from Web Corpora. PhD thesis, Masaryk University, Brno, Czech Republic.
- Quasthoff, U., M. Richter, and C. Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proc LREC*, 1799–1802. Genoa, Italy.
- Renouf, A., A. Kehoe, and J. Banerjee. 2006. WebCorp: an integrated system for web text search. *Language and Computers* 59(1): 47–67.
- Rundell, M. ed. 2002. *Macmillan English Dictionary for Advanced Learners*. Macmillan.
- Rundell, M. 2012. Stop the presses—the end of the printed dictionary. Macmillan Dictionary Blog, 5 Nov. <http://www.macmillandictionaryblog.com/bye-print-dictionary>.
- Rychlý, P. 2000. Korpusové manažery a ~ jejich efektivní implementace. Rychlý. PhD Thesis, Masaryk University, Brno, Czech Republic.
- Rychlý, P. 2007. Manatee/bonito—a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70. Masaryk University, Brno, Czech Republic.
- Sanseido. 2003, 2007. *The WISDOM English–Japanese Dictionary*. Sanseido.
- Schäfer, R., and F. Bildhauer. 2013. Web corpus construction. *Synthesis Lectures on Human Language Technologies* 6(4): 1–145.
- Scheible, S., R.J. Whitt, M. Durrell, and P. Bennett. 2011. A gold standard corpus of Early Modern German. In *Proc 5th Linguistic Annotation Workshop*, 124–128. Association for computational linguistics.
- Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*, eds. Baroni and Bernardini, 63–98. Bologna: Gedit.
- Srdanovic Erjavec, I., Erjavec, T., and Kilgarriff, A. 2008. A web corpus and word sketches for Japanese. *Information and Media Technologies*, 3(3).
- Suohomel, V., and J. Pomikálek. 2012. Efficient Web crawling for large text corpora. In *Proc Seventh Web as Corpus Workshop (WAC7)*, 39–43. Lyon, France.
- Thomas, J. 2014. *Discovering English with the Sketch Engine*. Print-on-demand. <http://ske.li/deske>.
- Tiedemann, J., and L. Nygaard. 2004. *The OPUS Corpus—parallel and free*. Lisbon: Proc LREC.
- Wild, K., A. Kilgarriff, and D. Tugwell. 2013. The Oxford Children's Corpus: using a Children's Corpus in Lexicography. *International Journal of Lexicography* 26(2): 190–218.