### Sketch Engine for English Language Learning

Vít Baisa Lexical Computing Ltd., UK and Masaryk Univ, Cz

Vít Suchomel Lexical Computing Ltd., UK and Masaryk Univ, Cz

Adam Kilgarriff Lexical Computing Ltd Brighton, UK

## {vit.baisa,vit.suchemel,adam.kilgarriff} @sketchengine.co.uk

There are many websites for language learners: wordreference.com, <sup>1</sup> Using English, <sup>2</sup> and many others. Some of them use corpus tools or corpus data such as Linguee, <sup>3</sup> Wordnik <sup>4</sup> and bab.la. <sup>5</sup> We introduce a novel free web service aimed at teachers and students of English which offers similar functions but is based on a specially prepared corpus suitable for language learners, using fully automated processing, offering the advantages that the corpus is very large – so can offer ample examples for even quite rare words and expressions. We call the service SkELL—Sketch Engine for Language Learning.<sup>6</sup>

SkELL offers three ways for exploring the language:

- Examples: for a given word or phrase up to 40 selected example sentences are shown
- Word sketch, showing typical collocates for the search term
- Similar words, visualized as a word cloud.

Examples (a simplified concordance) is a full-text search tool (see Figure 1).

Word sketches are useful for discovering collocates and for studying the contextual behaviour of words. Collocates of a word are words which occur frequently together with the word—they "co-locate" with the word. For a query, eg *language* (see Figure 2), SkELL will generate several lists containing collocates of the headword mouse. List headers describe what kind of collocates they contain. The collocates are shown in basic word forms, or 'lemmas'. By clicking on a collocate, a user can see a concordance with highlighted headwords and collocate (using red for the headword and green for the collocate).

The third tool shows words which are similar to a search word, in terms of 'sharing' the same collocates. They may be synonyms, near-synonyms or other related words. For a single word SkELL will return a list of up to forty of the words which are most similar. They are presented as a word cloud (Figure 3).

# 1.1 language learning

1.0 hits per million

- The highly engaging courses utilize progressive language learning methods.
- <sup>2</sup> These external characteristics may impact language learning opportunities.
- 3 Their language learning ability is very strong.
- 4 They are used very widely in language learning.
- 5 What is the best **language learning** software?
- <sup>6</sup> The development of **language learning** is thereby disrupted.
- 7 Language learning normally occurs most intensively during human childhood .
- <sup>8</sup> The same is true with children with **language learning** difficulties.
- 9 A broader approach to **language learning** than community language learning.
- <sup>10</sup> The importance of listening in **language learning** is often overlooked.

Figure 1: Examples for *language learning* 

#### The corpus

SkELL uses a large text collection (the 'SkELL corpus') gathered specially for the purpose. We had discovered previously that other collections of over a billion words at our disposal were all from the web, and contained too much spam to use for SKELL: it was critical not to show example sentences to learners if the sentences were not real English at all, but computer-generated junk. The SkELL corpus consists of spam-free texts from news, academic papers, Wikipedia, open-source (non)-fiction books, webpages, discussion forums, blogs etc. There are more than 60 million sentences in the corpus, and one and a half billion words. This volume of data provides a sufficient coverage of everyday, standard, formal and professional English language, even for mid-to-low frequency words and their collocations.

One of the biggest parts of SkELL corpus is English Wikipedia.<sup>7</sup> We included the 130,000 longest articles. Among the longest are articles on *South African labour law, History of Austria, Blockade of Germany:* there are many articles with geographical and historical texts.

Another substantial part consists of books from

<sup>&</sup>lt;sup>1</sup> http://www.wordreference.com

<sup>&</sup>lt;sup>2</sup> http://usingenglish.com

<sup>&</sup>lt;sup>3</sup> http://www.linguee.com/

<sup>&</sup>lt;sup>4</sup> https://www.wordnik.com

<sup>&</sup>lt;sup>5</sup> http://en.bab.la

<sup>&</sup>lt;sup>6</sup> While it was tempting to say the 'E' in the acronym should be for 'English', we decided against, as we envisage offering SkELL for other languages (SkELL-it, SKeLL-de, etc).

<sup>&</sup>lt;sup>7</sup> https://en.wikipedia.org

The Project Gutenberg<sup>8</sup>. The largest texts in the PG collection are *The Memoires of Casanova, The Bible (Douay-Rheims version), The King James Bible, Maupassant's Original short stories, Encyclopaedia Britannica.* 

We have also prepared two subsets from the enTenTen14, a large general web crawl (Jakubicek et al 2013). The 'White' (bigger) part contains only documents from web domains in www.dmoz.org or in the whitelist of www.urlblacklist.com, as the sites on these lists were known to contain only spam-free material. The 'Superwhite' (smaller) part contained documents from domains listed in the whitelist of www.urlblacklist.com—a subset of White (in case there is still some spam in the larger part taken from www.dmoz.org). The White part contained 1.6 billion tokens.

One part of the SkELL corpus has been built using WebBootCat (Pomikalek et al. 2006, an implementation of BootCaT (Baroni and Bernardini 2004)). This approach uses seed words to prepare queries for commercial search engines.<sup>9</sup> The pages from the search results are downloaded, cleaned and converted to plain text preserving basic structure tags. We assume the search results from the search engine are spam-free, because the search engines take great efforts not to return spam pages to users (wherever there are non-spam pages containing the search terms): BootCaT takes a 'free ride' on the anti-spam work done by the search engines. We have bootcatted approximately 100 million tokens.

We included all of the British National Corpus, as we know it to contain no spam.

The rest of the SkELL corpus consists of free news resources. Table 1 lists the sources used in the SkELL corpus.

Subcorpus	<b>Tokens</b> (= words + punctuation) millions	<b>Tokens used</b> millions
Wikipedia	1,600	500
Gutenberg	530	200
White	1,600	500
BootCatted	105	all
BNC	112	all
other resources	340	200

Table 1: Sources used for SkELL corpus

As the name says, SkELL builds on the Sketch Engine (Kilgarriff et al 2004), and the corpus was compiled using standard Sketch Engine procedures. We scored all sentences in the corpus using the GDEX tool for finding good dictionary examples (Kilgarriff et al 2008), and re-ordered the whole

corpus so it was sorted according to the score. This was a crucial part of the processing as it speeds up further querying. Instead of sorting good dictionary examples at runtime, all query results for concordance searches are shown in the sorted order without further work needing to be done.

The web interface is available at http://skell.sketchengine.co.uk. There is a version for mobile devices which is optimized for smaller screens and for touch interfaces, available at http://skellm.sketchengine.co.uk.

We have described a new tool which we believe will turn out to be very useful for both teachers and students of English. The processing chain is also ready to be used for other languages. The interface is also directly reusable for other languages, the only prerequisite is the preparation of the specialized corpus. We are gathering feedback from various users and will refine the corpus data and web interface accordingly in the future.

#### Acknowledgment

This work has been partly supported by the Ministry of Education of the Czech Republic within the LINDAT -Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

#### References

- Baroni, M., & Bernardini, S. (2004, May). BootCaT: Bootstrapping Corpora and Terms from the Web. In *LREC*.
- Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT* (pp. 247-252).
- Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen Corpus Family. In *Proc. Int. Conf. on Corpus Linguistics*.
- Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. In *Proceedings* of *EURALEX* (Vol. 6). Lorient, France. Pp 105-116.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008, July). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings* of *EURALEX* (Vol. 8).

<sup>&</sup>lt;sup>8</sup> https://www.gutenberg.org

<sup>&</sup>lt;sup>9</sup> We currently use the Bing search engine, http://www.bing.com

Search	Examples	Word sketch Similar w
1		
verbs with language as subject	adjectives with language	modifiers of language
learn	intelligible	programming
belong	Arabic	English
evolve	extinct	official
differ	akin	foreign
influence	ambiguous	native
tend	English	Romance
borrow	such	written
distinguish	unavailable	sign
accord	compulsory	Germanic
consist	identical	Slavic
emerge	similar	Indo-European
interpret	suitable	indigenous
lack	peculiar	modern
mean	distinct	European
undergo	Spanish	different
	verbs with language as subject learn belong evolve differ influence tend borrow distinguish accord consist emerge interpret lack mean	verbs with language as subjectadjectives with languagelearnintelligiblebelongArabicevolveextinctdifferakininfluenceambiguoustendEnglishborrowsuchdistinguishunavailableaccordcompulsoryconsistidenticalemergesimilarinterpretsuitablelackpeculiarmeandistinct

nouns modified by language	words and/or language
learner	<u>culture</u>
barrier	dialect
acquisition	literature
learning	custom
proficiency	religion
<u>skill</u>	mathematics
	Freihab

Figure 2: Word sketch for *language* 



picnic restaurant weddrg ride anlartainment

peolion

Figure 3: 'Similar words' word clouds for language and lunch. Size of the word in the word cloud represents similarity to the search word.