SkELL: Web Interface for English Language Learning

Vít Baisa and Vít Suchomel

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xbaisa,xsuchom2}@fi.muni.cz

Lexical Computing Ltd.
Brighton, United Kingdom
vit.{baisa,suchomel}@sketchengine.co.uk

Abstract. We present a new web interface for English language learning: SkELL. The name stands for Sketch Engine for Language Learning and is aimed at students and teachers of English language. We describe SkELL features and the processing of corpus data which is fundamental for SkELL: spam free, high quality texts from various domains including diverse text types covering majority of English language phenomena.

Keywords: Sketch Engine, concordance, thesaurus, word sketch, language learning, English language, corpus

1 Introduction

There are many websites for language learners: wordreference.com[1] and *Using English*[2] are just two of many. Some of them are using corpus tools or corpus data such as Linguee[3], Wordnik[4], bab.la[5]. They usually provide dictionary-like features: definitions and translation equivalents in selected languages. Some of them provide even examples from parallel corpora (Linguee).

We introduce here a new web interface aimed at teachers and students of English language which offers similar functions as above-mentioned tools but at the same time it is based on a specially processed corpus data suitable for the language learning purpose.

We call it SkELL: Sketch Engine for Language Learning. The Sketch Engine[6] is a state-of-the-art web-based tool for building, managing and exploring large

[1] http://www.wordreference.com/

[2] http://www.usingenglish.com/

[3] http://www.linguee.com/

[4] http://www.wordnik.com/

[5] http://en.bab.la/

[6] http://www.sketchengine.co.uk

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 63–70, 2014. © NLP Consulting 2014

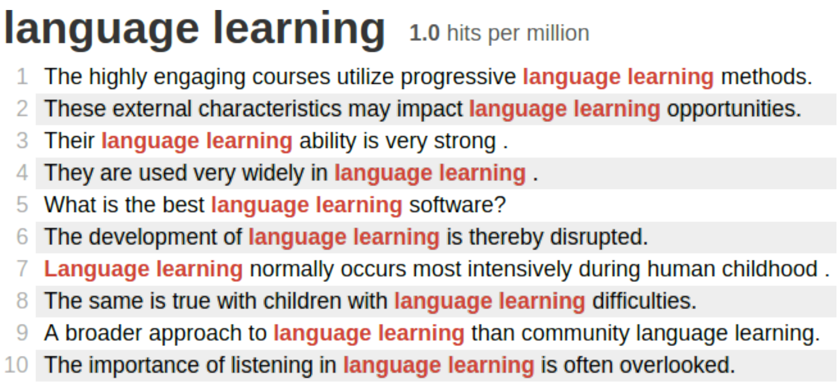
text collections in dozens of languages. SkELL is derived from the Sketch Engine and the data SkELL relies upon (SkELL corpus) is built using the very same tools as Sketch Engine uses: web crawler SpiderLing [1], tokeniser unitok.py [2] and TreeTagger [3]. Also, it uses a technique for scoring sentences according their appropriateness for using as example sentences in learners' dictionaries, GDEX [4].

2 Features of SkELL

SkELL features offer three ways for exploring the SkELL corpus. The first is the *concordance*: for a given word or phrase, it will return up to 40 example sentences. The second is the *word sketch* through which typical collocates for a given word can be discovered. And the third is *similar words* (thesaurus) which lists words that are similar to, though not necessarily synonymous with, a search word. The similar words are visualized with a word cloud. The web interface is optimized for mobile and touch devices.

SkELL features are built upon Bonito corpus manager [5] features. Bonito provides many standard functions as many other corpus managers: concordancing, word list generating, context statistics and also some advanced features like distributional thesaurus [6] and word sketches [7]. We have chosen these three: 1) concordance, 2) word sketch and 3) thesaurus (similar words).

2.1 Concordance (or examples)



language learning 1.0 hits per million

1 The highly engaging courses utilize progressive **language learning** methods.
2 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?
6 The development of **language learning** is thereby disrupted.
7 **Language learning** normally occurs most intensively during human childhood .
8 The same is true with children with **language learning** difficulties.
9 A broader approach to **language learning** than community language learning.
10 The importance of listening in **language learning** is often overlooked.

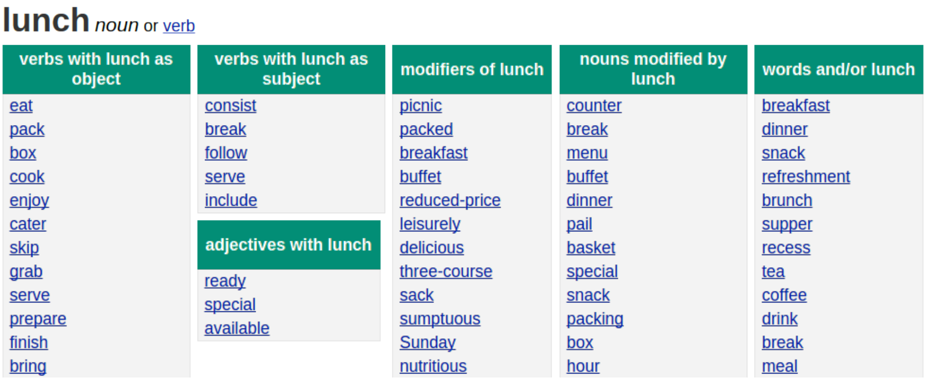
Fig. 1: Example of concordance for *language learning* phrase

Concordance offers a powerful full-text search tool. For a word or a phrase it returns up to 40 example sentences featuring the query words. Concordance feature is useful for discovering how words behave in English.

The search is case insensitive, i.e. it will yield the same results for *rutherford* and *Rutherford*. Moreover the results may contain the query (a word or a phrase) in a derived word form. For *mouse* (lemma) it will find sentences also with *mice*. For *mice* the result will contain a different set of sentences: only *mice* occurrences.

It is not necessary for users to specify part of speech (PoS, e.g. noun, verb, adjective, preposition, adverb etc.) of the search term is not necessary. If you search for *book*, it will give sentences with *book* as a verb and as a noun and both in various word forms (*booking*, *booked*, *books*).

2.2 Word sketch (or word profile)



lunch *noun* or verb

verbs with lunch as object	verbs with lunch as subject	modifiers of lunch	nouns modified by lunch	words and/or lunch
eat	consist	picnic	counter	breakfast
pack	break	packed	break	dinner
box	follow	breakfast	menu	snack
cook	serve	buffet	buffet	refreshment
enjoy	include	reduced-price	dinner	brunch
cater	**adjectives with lunch**	leisurely	pail	supper
skip		delicious	basket	recess
grab	ready	three-course	special	tea
serve	special	sack	snack	coffee
prepare	available	sumptuous	packing	drink
finish		Sunday	box	break
bring		nutritious	hour	meal

Fig. 2: Example of word sketch for *lunch*.

Word sketch function is very useful for discovering collocates and for studying contextual behaviour of words. Collocates of a word are words which occur frequently together with the word—they "co-locate" with the word. See [7] for more info.

For query *mouse*, SkELL will generate several tables containing collocates of the headword *mouse*. Table headers describe what kind of collocates (always in basic word form) they contain.

By clicking on a collocate, a concordance with highlighted headwords and collocates is shown. This way it can be seen how the two collocates together are usually used in English language.

By default, the most frequent PoS is shown in Word Sketches. If a word (*book*, *fast*, *key*, ...) can have more than one PoS, the alternative links are shown next to the headword (see in Figure 2).

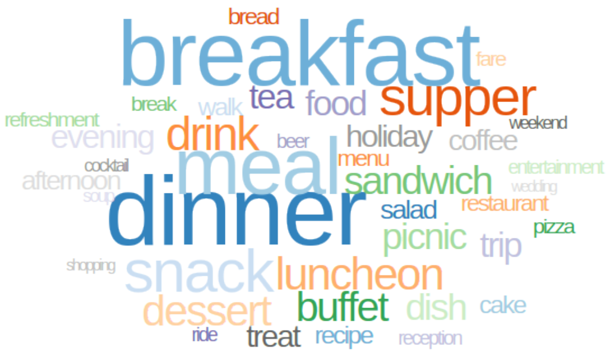


Fig. 3: Example of similar words in word cloud for *lunch*.

2.3 Similar words (or thesaurus)

The third functions serves for finding words which are similar (not only synonyms) to a search word. For a word (multi word queries are not supported yet) SkELL will return list of up to 40 most similar words visualized using wordcloud. The wordcloud is generated using D3.js library[7] and a wordcloud plugin[8].

As in Word Sketch if a word can have more than one PoS, the links to alternative results are provided.

3 SkELL corpus

SkELL is using a large text collection—SkELL corpus—gathered specially for the purpose of English language learning. It consists of texts from news, academic papers, Wikipedia articles, open-source (non)-fiction books, webpages, discussion forums, blogs etc. There are more than 60 million sentences in SkELL corpus and more than one billion words in total. This amount of textual data provides a sufficient coverage of everyday, standard, formal and professional English language.

In the following subsections we describe the most important data resources which have been used in building the SkELL corpus and the processing of the data.

3.1 English Wikipedia

One of the largest parts of SkELL corpus is English Wikipedia.[9] We have used Wikipedia XML dump[10] from October 2014. The XML file has been converted

[7] `http://d3js.org/`

[8] `http://www.jasondavies.com/wordcloud/about/`

[9] `https://en.wikipedia.org`

[10] `https://dumps.wikimedia.org/`

to plain text preserving only a few structure tags (documents, headings, paragraphs). using slightly modified script `WikiExtractor.py`[11]. We have filtered thousands of articles which are supposed to not contain fluent English text, e.g. articles named "List of ..." and then sorted all articles according their length and took top 130,000 articles. Among the longest articles there were e.g.: *South African labour law*, *History of Austria*, *Blockade of Germany*, ... It is clear that there are many articles from geographical and historical domains.

3.2 Project Gutenberg

The Project Gutenberg[12] (PG) focuses on gathering public domain texts in many languages. The majority of texts is in English. We have downloaded all English texts using *wget*[13]. and converted the HTML files to plain text.

The largest texts in English PG collection are *The Memoires of Casanova*, *The Bible (Douay-Rheims version)*, *The King James Bible*, *Maupassant's Original short stories*, *Encyclopaedia Britannica*, etc.

3.3 English web corpus, enTenTen14

We have prepared two subsets from the enTenTen14 [8] which has been crawled in 2014. The *White* (bigger) part contains only documents from web domains in `dmoz.org` or in the whitelist of `urlblacklist.com`. The *Superwhite* (smaller) containing documents domains listed in the whitelist of `urlblacklist.com` – a subset of *White* (in case there is still some spam in the larger part taken from `dmoz.org`)

Categories from the following list are allowed categories from `dmoz.org` in *Superwhite* part: 1) blog: journal/diary websites, 2) childcare: sites to do with childcare, 3) culinary: sites about cooking, 4) entertainment: sites that promote movies, books, magazine, humor, 5) games: game related sites, 6) gardening: gardening sites, 7) government: military and schools etc., 8) homerepair: sites about home repair, 9) hygiene: sites about hygiene and other personal grooming related stuff, 10) medical: medical websites, 11) news: news sites, 12) pets: pet sites, 13) radio: non-news related radio and television, 14) religion: sites promoting religion, 15) sportnews: sport news sites, 16) sports: all sport sites, 17) vacation: sites about going on holiday, 18) weather: weather news sites and weather related and 19) whitelist: sites specifically 100% suitable for kids. Finally we have decided to include the whole *White* part. It contained 1.6 billion tokens.

[11] `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

[12] `http://www.gutenberg.org/`

[13] `http://www.gutenberg.org/wiki/Gutenberg:Information_About_Robot_Access_to_our_Pages`

3.4 WebBootCat corpus

One part of the SkELL corpus has been built using WebBootCat [9]. This approach uses seed words to prepare queries for a search engine. 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. We have run the tool several times with general English words as seed words yielding approximately 100 million tokens.

3.5 Other resources

The whole British National Corpus [10] has been also included. The rest of the SkELL corpus consists of free news sources. The Table 1 contains all the sources used in SkELL corpus.

Table 1: Sources used in SkELL corpus

Subcorpus	tokens	used
Wiki	1.6 G	500 M
Gutenberg	530 M	200 M
White	1.6 G	500 M
WebBootCated	105 M	all
BNC	112 M	all
other sources	344 M	200 M

3.6 Processing the data

When all the subparts have been gathered and pre-cleaned (we have removed all structures except sentences), we have run it through our standard processing pipe:

1. normalization: quotes, interpunction normalization,
2. tokenization: we have used `unitok.py`,[14]
3. TreeTagger[15] for English,
4. deduplication on sentence level: `onion` tool[16] has been used.

The corpus was then compiled using manatee indexing library [5]. Then we have scored all sentences in the corpus using GDEX tool [4] for finding good dictionary examples (sentences) for query results. All sentences in the corpus were sorted according to the score and saved in this order. This is a crucial part of the processing as it speeds up further querying of the corpus. Instead of

[14] https://corpus.tools

[15] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[16] http://nlp.fi.muni.cz/projects/onion/

sorting good dictionary examples on-the-fly (which is used in Sketch Engine), all query results for concordance searches are presorted in the source vertical file.

The standard GDEX definition file used for English has been only slightly changed to prefer sentences with more frequent words, filtering out effectively all sentences with special terminology, typos and rare words (rare names). By default, short sentences are preferred, sentences containing inappropriate or spam words are scored lower.

The word sketch grammar required for computing word sketches has been modified: it contains only a few grammar rules with self-explanatory names.

3.7 Versioning and referencing

Since SkELL corpus may be changed in the future (further cleaned, refined, updated), all references to particular results of SkELL should be accompanied by the current version. The web interface may also be changed occasionally. That is why at the bottom of SkELL page, there is a version in this format: version1-version2. The first corresponds to a version of the web interface and the second to a version of SkELL corpus.

4 Conclusions and future work

The web interface is available at `http://skell.sketchengine.co.uk`. The version for mobile devices which is optimized for smaller screens and for touch interfaces is available at `http://skellm.sketchengine.co.uk`. If you access the former link from a mobile device it should be detected and redirected to the mobile version automatically.

We have described a new tool which we belive will turn out to be very useful for both teachers and students of English. The processing chain is ready to be used also for other languages. The interface is also directly reusable for other languages, the only prerequisite is the preparation of the corpus.

We are gathering feedback from various users and will refine the corpus data according it. In the future we plan these updates to SkELL:

1. to create a special grammar relation for English phrasal verbs,
2. to combine examples for collocations from word sketch collocates to build a more representative concordance for a given word,
3. to update the corpus with the newest text resources to provide examples for the newest trending words and neologisms,
4. to analyse access logs of SkELL and put favourite searches to the main page,
5. to provide commonest string [11] for word sketch collocates,
6. to allow multi word sketches which have been already introduced in [11],
7. to implement necessary methods for multi word thesaurus and
8. to build SkELL also for other languages (Russian, Czech, German and other).

Acknowledgement This work was partially supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

References

1. Suchomel, V., Pomikálek, J., et al.: Efficient web crawling for large text corpora. In: Proceedings of the seventh Web as Corpus Workshop (WAC7). (2012) 39–43
2. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text Tokenisation Using unitok. In Horák, A., Rychlý, P., eds.: 8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2014) 71–75
3. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods in language processing. Volume 12., Manchester, UK (1994) 44–49
4. Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P.: Gdex: Automatically finding good dictionary examples in a corpus. In: Proceedings of EURALEX. Volume 8. (2008)
5. Rychlý, P.: Manatee/bonito–a modular corpus manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing, within MU: Faculty of Informatics Further information (2007) 65–70
6. Rychlý, P., Kilgarriff, A.: An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics (2007) 41–44
7. Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D.: The Sketch Engine. Information Technology **105** (2004)
8. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V., et al.: The tenten corpus family. In: Proc. Int. Conf. on Corpus Linguistics. (2013)
9. Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P., et al.: Webbootcat: instant domain-specific corpora to support human translators. In: Proceedings of EAMT. (2006)
10. Leech, G.: 100 million words of english: the british national corpus (bnc). Language Research **28**(1) (1992) 1–13
11. Kilgarriff, A., Rychly, P., Kovář, V., Baisa, V.: Finding multiwords of more than two words. Proceedings of EURALEX2012 (2012)