Title: Sketch Engine

URL: https://en.wikipedia.org/wiki/Sketch_Engine

PageID: 43563308

Categories: Category:Applied linguistics, Category:Computational linguistics, Category:Corpus linguistics, Category:Data mining and machine learning software, Category:Database management systems, Category:Lexicography, Category:Linguistic research, Category:Natural language processing, Category:Text analysis, Category:Text mining

---

Sketch Engine is a corpus manager and text analysis software developed by Lexical Computing since 2003. Its purpose is to enable people studying language behaviour (lexicographers , researchers in corpus linguistics , translators or language learners) to search large text collections according to complex and linguistically motivated queries. Sketch Engine gained its name after one of the key features, word sketches : one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour. [2] Currently, it supports and provides corpora in over 100 languages. [3]

History of development

Sketch Engine is a product of Lexical Computing, a company founded in 2003 by the lexicographer and research scientist Adam Kilgarriff . [4] He started a collaboration with Pavel Rychlý, a computer scientist working at the Natural Language Processing Centre, Masaryk University , [5] and the developer of Manatee and Bonito (two major parts of the software suite). Kilgarriff also introduced the concept of word sketches .

Since then, Sketch Engine has been commercial software, however, all the core features of Manatee and Bonito that were developed by 2003 (and extended since then) are freely available under the GPL license within the NoSketch Engine suite. [6]

Features

A list of tools available in Sketch Engine:

Word sketches – a one-page automatic derived summary of a word's grammatical and collocational behaviour

Word sketch difference – compares and contrasts two words by analysing their collocations

Distributional thesaurus – automated thesaurus for finding words with similar meaning or appearing in the same/similar context

Concordance search – finds occurrences of a word form, lemma , phrase, tag or complex structure

Collocation search – word co-occurrence analysis displaying the most frequent words (for a search word) which can be regarded as collocation candidates

Word lists – generates frequency lists which can be filtered with complex criteria

n-grams – generates frequency lists of multi-word expressions

Terminology / Keyword extraction (both monolingual and bilingual) – automatic extraction of key words and multi-word terms from texts (based on frequency count and linguistic criteria)

Diachronic analysis (Trends) [7] – detecting words which undergo changes in the frequency of use in time (show trending words)

Corpus building and management – create corpora from the Web or uploaded texts including part-of-speech tagging and lemmatization which can be used as data mining software

Parallel corpus (bilingual) facilities – looking up translation examples (EUR-Lex corpus, Europarl corpus , OPUS corpus, etc.) or building a parallel corpus from own aligned texts

Text type analysis – statistics of metadata in the corpus

Keywords and terminology extraction

Sketch Engine can perform automatic term extraction by identifying words typical of a particular corpus, document, or text. Single words and multi-word units can be extracted from monolingual or bilingual texts. The terminology extraction feature provides a list of relevant terms based on comparison with a large corpus of general language. This functionality is also available as a separate service called OneClick Terms with a dedicated interface. [8]

SKELL

A free web service based on Sketch Engine and aimed at language learners and teachers is SKELL (formerly SkELL). It exploits Sketch Engine's proprietary GDEX (Good Dictionary Examples) scoring function to provide authentic example sentences for specific target words. Results are drawn from a special corpus of high-quality texts covering everyday, standard, formal, and professional language and displayed as a concordance . SKELL also includes simplified versions of Sketch Engine's word sketch and thesaurus functions. [9]

It has been suggested that SKELL can be used, for instance, to help students understand the meaning and/or usage of a word or phrase; to help teachers wanting to use example sentences in a class; to discover and explore collocates ; to create gap-fill exercises ; to teach various kinds of homonyms and polysemous words . [10] [11] SKELL was first presented in 2014, when only English was supported. [9] Later, support was added for Russian , [12] Czech , [13] German , [14] Italian [15] and Estonian . [16]

List of text corpora

Sketch Engine provides access to more than 800 text corpora. There are monolingual as well as multilingual corpora of different sizes (from one thousand words up to 85 billion words) and various sources (e.g. web, books, subtitles, legal documents). The list of corpora includes British National Corpus , Brown Corpus , Cambridge Academic English Corpus and Cambridge Learner Corpus, CHILDES corpora of child language, OpenSubtitles (a set of 60 parallel corpora), 24 multilingual corpora of EUR-Lex documents, the TenTen Corpus Family (multi-billion web corpora), and Trends corpora (monitor corpora with daily updates). [17]

Architecture

Sketch Engine consists of three main components: an underlying database management system called Manatee, a web interface search front-end called Bonito, and a web interface for corpus building and management called Corpus Architect. [18]

Manatee

Manatee is a database management system specifically devised for effective indexing of large text corpora. It is based on the idea of inverted indexing (keeping an index of all positions of a given word in the text). It has been used to index text corpora comprising tens of billions of words. [19]

Searching corpora indexed by Manatee is performed by formulating queries in the Corpus Query Language (CQL). [20]

Manatee is written in C++ and offers an API for a number of other programming languages including Python , Java , Perl and Ruby . Recently, it was rewritten into Go for faster processing of corpus queries. [21]

Bonito

Bonito is a web interface for Manatee providing access to corpus search. In the client–server model , Manatee is the server and Bonito plays the client part. It is written in Python . [18]

Corpus Architect

Corpus Architect is a web interface providing corpus building and management features. It is also written in Python .

Applications

Sketch Engine has been used by major British and other publishing houses for producing dictionaries such as Macmillan English Dictionary , Dictionnaires Le Robert , Oxford University Press or Shogakukan . Four of United Kingdom's five biggest dictionary publishers use Sketch Engine. [22]

References

Further reading

Thomas, James (March 2016). Discovering English with Sketch Engine : a corpus-based approach to language exploration. Workbook and glossary . Brno: Versatile. ISBN 9788026095798 .

External links

Sketch Engine website

List of corpora available in Sketch Engine

OneClick terms – online term extractor with term extraction technology from Sketch Engine

SKELL – Sketch Engine for language learning

v

t

e

American National Corpus

Bank of English

Bergen Corpus of London Teenage Language

British National Corpus

Brown Corpus

Buckeye Corpus

Cambridge English Corpus

Corpus of Contemporary American English

Enron Corpus

EnTenTen

International Corpus of English

Lancaster-Oslo-Bergen Corpus

Oxford English Corpus

PropBank

Spoken English Corpus

Switchboard Telephone Speech Corpus

TIMIT

VerbNet

Wellington Corpus of Spoken New Zealand English

Bijankhan Corpus

CHILDES

CorCenCC National Corpus of Contemporary Welsh

Croatian Language Corpus
Croatian National Corpus
Czech National Corpus
Europarl Corpus
German Reference Corpus
Hamshahri Corpus
National Corpus of Polish
Neo-Assyrian Text Corpus Project
Persian Speech Corpus
Quranic Arabic Corpus
Russian National Corpus
Somali Corpus
Scottish Corpus of Texts and Speech
Slovenian National Corpus
TalkBank
Tatoeba
Tekstaro de Esperanto
TenTen Corpus Family
Thesaurus Linguae Graecae
BNC consortium
COBUILD
Sketch Engine