

# Interactive Visualizations of Corpus Data in Sketch Engine

Lucia Kocincová<sup>†</sup>  
xkocinc@fi.muni.cz

Vít Baisa<sup>‡†</sup> and Miloš Jakubíček<sup>‡†</sup> and Vojtěch Kovář<sup>‡†</sup>  
name.surname@sketchengine.co.uk

<sup>†</sup>NLP Centre, Faculty of Informatics  
Masaryk University, Brno, Czech Republic  
<sup>‡</sup>Lexical Computing, Brighton, United Kingdom

## Abstract

Automatic analysis of large text corpora produces large amounts of figures as result of various functions. These provide empirical evidence for a research hypothesis or serve in numerous practical applications of natural language processing. Usually, the results are presented in the form of tables containing raw data to be interpreted by domain experts. This paper describes an ongoing work on new visualizations and user interface enhancements in Sketch Engine corpus management system which aim at easing the interpretation of the data for both novice users and language professionals.

## 1 Introduction

Analyses of textual data deal with the issue of choosing a suitable representation for its results. While this factor is often neglected and most attention is being paid to the performance of the analytic functions (where the problem might be simply seen as continuation of Aristotle’s *form vs. matter* debate, with matter being absolutely predominant in science), there is no doubt that the representation heavily influences how data are perceived and can significantly help or harm correct understanding of the results (Meirelles, 2013).

This becomes even more appealing where the underlying analytic sample does not possess uniform distribution—like language which usually follows Zipf’s distribution (Zipf, 1949). Not only “ordinary” language users but sometimes even language experts tend to underestimate the impact of such heavily skewed distributions like the Zipfian one, henceforth drawing invalid conclusions from the analyses they carry out.

In this paper we describe an ongoing work on implementing new visualization options for language data analysis within Sketch Engine corpus management system (Kilgarriff et al., 2014). Sketch Engine has a large variety of users ranging from language learners, students and researchers in linguistics, lexicographers, translators and terminologists or data scientists in diverse domains.

The presented enhancements to the user interface are implemented with the hope that they will not only provide a more graphically appealing and easier way how to understand the data for novice users, but also speed up the daily work carried out by language experts (e.g. lexicographers).

## 2 Sketch Engine

Sketch Engine is a leading corpus management system useful for discovering how words behave in different contexts. It has a wide range of analytic functions dealing with billion-word corpora (see e.g. (Pomikálek et al., 2012; Jakubíček et al., 2013)). In this paper we focus on the visualization of two core functions that leverage the principles of distributional semantics—word sketches and a distributional thesaurus.

### 2.1 Word Sketches

A word sketch is a one-page summary of a word’s collocational behavior according to particular grammar relations. It is usually computed by evaluating a large number of corpus queries (Jakubíček et al., 2010) performing shallow parsing or by using some existing parser to do this task so as to retrieve a large number of collocation candidates which are then sorted using a lexicographic association score (see (Rychlý, 2008; Evert, 2005)).

A word sketch is currently presented in a table

**strategy** (noun)  
English Corpus for SkELL freq = [117,755](#) (79.07 per million)

modifiers of strategy	95,790	2.00	verbs with strategy as object	41,483	2.30	words and/or strategy	22,751	1.10	verbs with strategy as subject	18,846	1.40
marketing	<a href="#">2,143</a>	8.64	implement	<a href="#">1,271</a>	8.37	tactic	<a href="#">910</a>	8.70	backfire	<a href="#">27</a>	5.41
effective	<a href="#">1,113</a>	7.14	cope	<a href="#">468</a>	8.31	objective	<a href="#">196</a>	6.04	aim	<a href="#">87</a>	5.34
overall	<a href="#">791</a>	7.04	devise	<a href="#">431</a>	7.85	marketing	<a href="#">205</a>	6.02	involve	<a href="#">362</a>	5.08
long-term	<a href="#">610</a>	7.00	develop	<a href="#">2,848</a>	7.61	planning	<a href="#">191</a>	5.81	focus	<a href="#">126</a>	5.00
investment	<a href="#">789</a>	6.75	adopt	<a href="#">995</a>	7.51	policy	<a href="#">650</a>	5.56	depend	<a href="#">87</a>	4.94
exit	<a href="#">400</a>	6.74	formulate	<a href="#">271</a>	7.28	technique	<a href="#">314</a>	5.54	fail	<a href="#">133</a>	4.80
management	<a href="#">1,083</a>	6.71	pursue	<a href="#">528</a>	7.23	tip	<a href="#">110</a>	5.43	target	<a href="#">46</a>	4.41
prevention	<a href="#">372</a>	6.62	employ	<a href="#">678</a>	7.18	vision	<a href="#">142</a>	5.40	work	<a href="#">372</a>	4.39
comprehensive	<a href="#">431</a>	6.59	outline	<a href="#">272</a>	6.92	plan	<a href="#">480</a>	5.10	rely	<a href="#">35</a>	4.26
alternative	<a href="#">427</a>	6.59	execute	<a href="#">291</a>	6.69	prevention	<a href="#">53</a>	5.08	prove	<a href="#">93</a>	4.06

Figure 1: Word sketches for lemma *strategy*.

where each column contains one grammar relation with collocates sorted by their score (see Figure 1).

## 2.2 Thesaurus

On top of the word sketches, Sketch Engine computes a distributional thesaurus. Normally a distributional thesaurus tackles the issue of finding similar words by asking the following question: given a word, what are the words that occur in the same context? In Sketch Engine this question is answered by finding words that share the same collocates in the same grammar relations in word sketches.

A thesaurus is simply a list of words accompanied with their frequency and a similarity score.

## 3 Thesaurus Visualization

The thesaurus already provides a visualization of the words in the form of a word cloud (Figure 2), however the score is not represented in the thesaurus in any way, so the potential of the data is not used to its full extent.

The main objective of the presented visualization in Figure 6 is to display all thesaurus attributes (textual string, its frequency and similarity and score) of each lemma in a meaningful but also a clear way, which was achieved by mapping these two values to multiple attributes. The core of the visualization is a lemma, the respective words are placed around it according to the score value—the higher the score is, the closer a word is to the center.

The score values are normalized into the  $[0, 1]$  range, therefore a comparison of two different word lists is possible. However, a user evaluates each word list independently, so a fixed score axis could lead into misunderstanding in cases where

score value is relatively low. This fact was taken into account when designing the visualization, so the score axis is adaptive – its boundary values are always taken from currently selected data, therefore the user can always clearly indicate the most similar words of the current lemma.

Score in the visualization is also mapped to a colour of the circle behind the word which simplifies comparison of two close words. The user can modify the colour range of score by choosing another appropriate colour from control panel, so the visual output can be adjusted.

Frequency is mapped to the size of a circle and also to the font size of a word, which can be disabled to avoid confusion when comparing short and long words.

## 4 Word Sketch Visualization

Word sketches are technically very similar in terms of data types—thesaurus is a word list with score and frequency and word sketches are multiple word lists with the same attributes. Therefore, the principles in graphical representations may remain the same—assuming also that a consistent visualization across different system function makes its perception easier.

Score and frequency values are mapped to the same graphical elements as in thesaurus except the colour of circles because in word sketches, distinct colours are used for different grammar relations – as can be perceived from Figure 3. Therefore a score on different radius around the center is mapped to circle transparency so the words with the highest scores pop out from the center of the picture.

Example of word sketch visualization can be seen in Figure 4.

**strategy** *(noun)*  
 English Corpus for SkELL freq = [117,755](#) (79.07 per million)

Lemma	Score	Freq
<a href="#">policy</a>	0.591	279,783
<a href="#">approach</a>	0.580	166,350
<a href="#">plan</a>	0.577	311,681
<a href="#">development</a>	0.564	352,390
<a href="#">method</a>	0.556	224,044
<a href="#">practice</a>	0.545	232,839
<a href="#">process</a>	0.537	373,993
<a href="#">technique</a>	0.535	113,354
<a href="#">management</a>	0.533	171,066
<a href="#">effort</a>	0.531	251,057
<a href="#">activity</a>	0.530	285,931
<a href="#">research</a>	0.522	284,144
<a href="#">program</a>	0.520	432,606
<a href="#">technology</a>	0.509	202,056
<a href="#">solution</a>	0.507	143,472
<a href="#">project</a>	0.506	325,209
<a href="#">concept</a>	0.505	130,715
<a href="#">measure</a>	0.497	132,326
<a href="#">design</a>	0.494	251,627
<a href="#">initiative</a>	0.492	55,995



Figure 2: Thesaurus for lemma *strategy*.

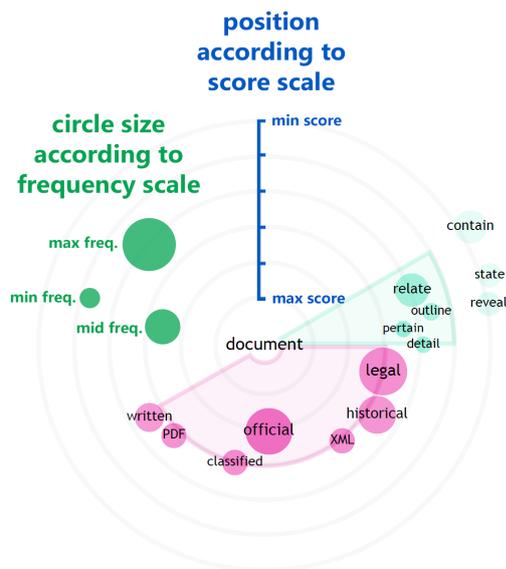


Figure 3: Mapping of graphical elements.

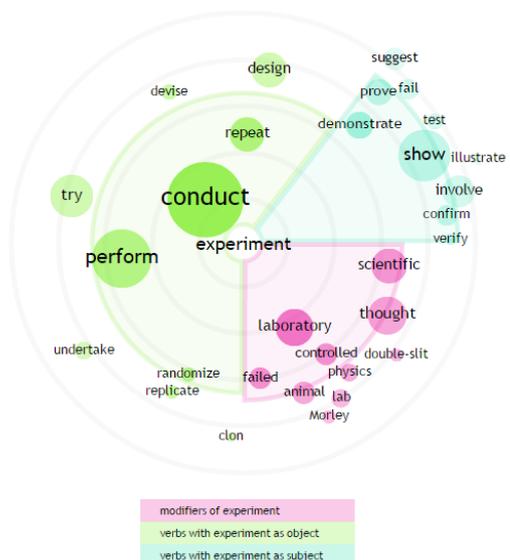


Figure 4: Word sketches of *experiment*.

## 5 Interactivity of Visualizations

Each visualization is interactive thus the exploration of relationships among words is easier. All interaction controls are grouped in a panel located in the right side of the page. The panel includes as many options as possible, however the number of options will be reduced after a user testing.

The exact values of a word's frequency and score are not left out entirely, they can be retrieved on demand by hovering over a particular word. For an approximate evaluation of values which is

mostly used by users, boundaries with labels are located in a legend. It is automatically updated when data changes so the user is always aware of the applied scale.

The interfaces and their parts are described in Figure 5.

## 6 Implementation

The visualizations introduced in this paper were implemented using JavaScript, jQuery and D3 library. D3 helps to focus on the graphical output

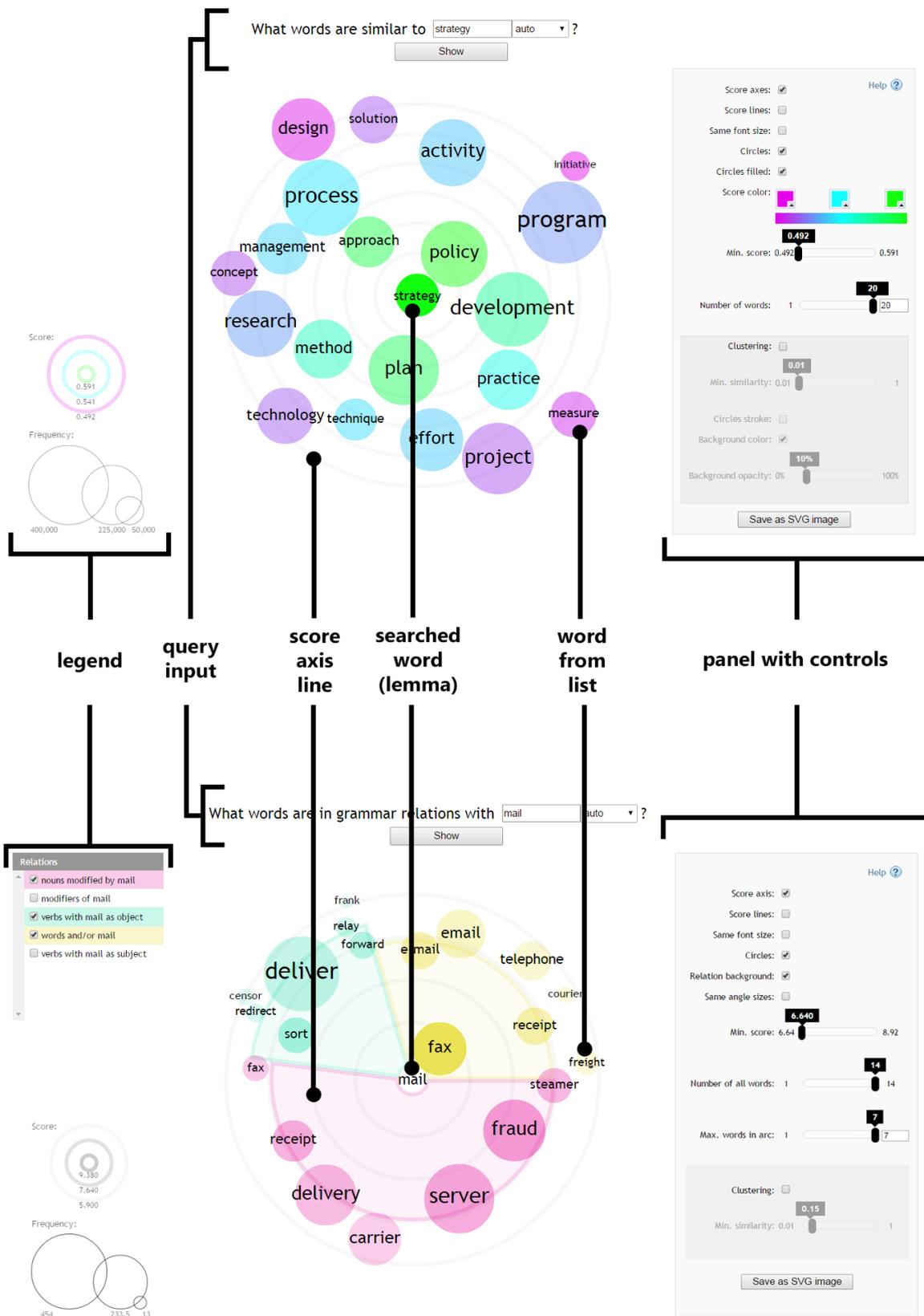


Figure 5: Visualization interface of thesaurus (top) and word sketches (bottom).

and its transformations and allows more effective development of interactive visualizations. (Bostock et al., 2011)

The input for scripts generating the visual output is a JSON object retrieved from Sketch Engine. Data boundaries and other properties are calculated to set up scales correctly.

An assignment of coordinates for each word is made afterwards according to score values. The new position has to satisfy two conditions—the word’s bounding box cannot overlap with other bounding boxes in the area and also the circles cannot overlap. If these conditions are not met all scales are contracted and the positions are recalculated. If it is not possible to meet the given restrictions, for example too many words are being requested for output, the limitations are dropped and positions are calculated without any restrictions and it is upon the user to filter the output.

This behavior ensures that a visualization is always rendered and the user’s requirements for data exploration are not limited.

The exact positions of words in a given score radius are currently random, but mapping of another attribute is possible in the future.

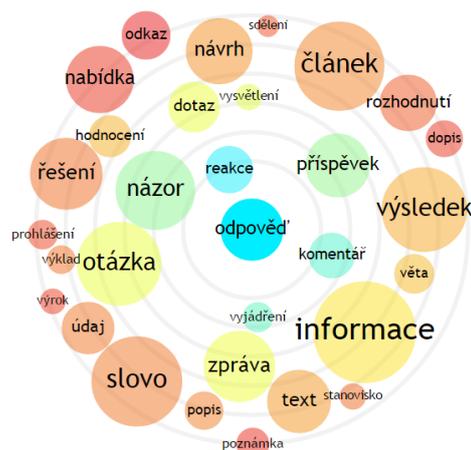
In the visualization of thesaurus the whole area around the center is covered with words. In the word sketches the whole area is divided into grammar relations, each relation is assigned an arc whose area is calculated as the sum of frequencies of words that belong to the given relation. The length of an arc represents the average score of displayed words.

The presented visualizations don’t evaluate or modify the words as strings—the algorithms work with them as elements—they are therefore language-independent and can be used with any corpora available in Sketch Engine as can be seen in Figure 6.

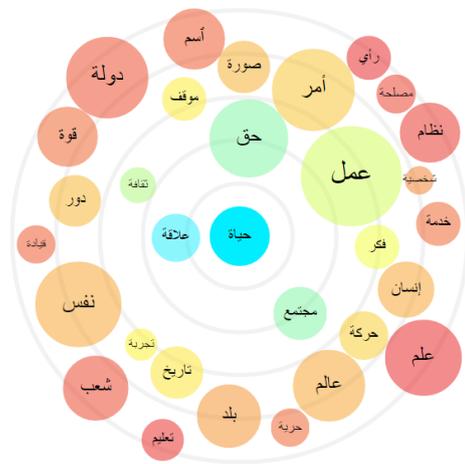
Graphics generated from the system can be also downloaded for further use.

## 7 Conclusions and Future Work

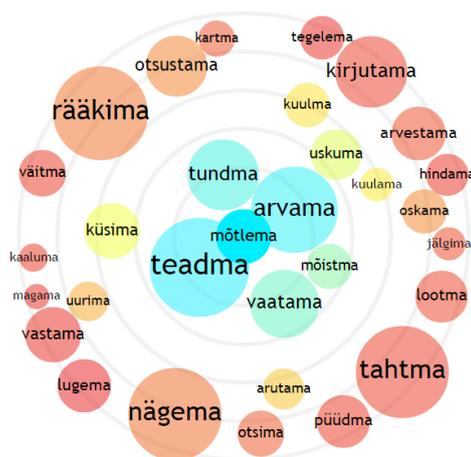
In this paper we have presented a number of visualization enhancements that are soon to appear in Sketch Engine (including its open source variant NoSketch Engine). The main added value of these visualizations is that users can immediately see differences between data values which can lead to faster interpretation of results and faster decisions. We are confident that further development of such



(a) Czech corpus



(b) Arabic corpus



(c) Estonian corpus

Figure 6: Thesaurus generated from different corpora.

enhancements is necessary to facilitate better understanding of the underlying corpus data especially in the context of (“big data”).

Evaluation testing with differently experienced users of Sketch Engine is currently in progress and so far shows that the visualizations are mostly valuable for new and intermediate users. According to the feedback from users we also plan A/B testing to verify new improved versions of the presented visualizations. In the future further features of Sketch Engine will be subject to visualization enhancements as well (e. g. corpus metadata overview).

## Acknowledgments

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarín project LM2010013. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under ProjectContract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme “Projects of Large Infrastructure for Research, Development, and Innovations” (LM2010005) is appreciated.

## References

- Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D<sup>3</sup> Data-Driven Documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309.
- Stefan Evert. 2005. *The statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart, Holzgartenstr. 16, 70174 Stuttgart.
- Miloš Jakubíček, Adam Kilgarriff, Diana McCarthy, and Pavel Rychlý. 2010. Fast Syntactic Searching in Very Large Corpora for Many Languages. *PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 741–747.
- Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen Corpus Family. *International Conference on Corpus Linguistics, Lancaster*.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel

Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1:7–36.

Isabel Meirelles. 2013. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport publishers.

Jan Pomikálek, Pavel Rychlý, and Miloš Jakubíček. 2012. Building a 70 Billion Word Corpus of English from ClueWeb. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, pages 502–506.

Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.