

The Role of Sketch Engine in Multiple Types of Corpora



Sharipova Aziza Abdumanapovna

Abstract: A large number of dictionary projects use Sketch Engine as it is the most renowned corpus management and corpus query tool. This paper sheds light on the significant role Sketch Engine plays in relation to different types of corpora. The software's features that support the creation of multilingual dictionaries and lexicography are also discussed. The software keeps on updating its old features to facilitate the users, and also adds entirely new features every now and then.

Keywords: Bilingual Corpus, Corpora, Multilingual Corpus, Sketch Engine

I. INTRODUCTION

In 2003, Adam Kilgarriff created the leading text analysis and corpus manager software "Sketch Engine" (Kilgarriff et al. 2014). It is widely used in well-known international institutions, such as Macmillan and Oxford University, particularly for their large learners' dictionary projects.

According to Rychlý (2007), the fast assessment of queries over gigantic corpora is made possible through "Manatee" which is a specialized data base engine that provides basis for the system. The corpus size is not confined by Sketch Engine; so far, an 80 billion words corpus has been the largest corpus to be tested (Kilgarriff, A., Jakubiček, M., Kovář, V., Rychlý, P. and Suchomel, V. 2014). In order to present the whole system of Sketch Engine as a web application, it uses Bonito; a web-based graphical user interface since 2004. The Corpus Architect is a corpus building and management module which was introduced by Sketch Engine in 2010. This permits the users to simply create user corpora from the web or using their own data. The language independency of all system parts is guaranteed. The corpora tend to have the corresponding metadata when users are searching for these comparable corpora. An example of this corpora is the CHILDES corpora in the Sketch Engine, and the different corpora found in the Wikipedia. (Abdumanapovna, S.A. 2018) Word Sketch is one of the eminent features of Sketch Engine. It derives statistical collocation behavior of a specific word from the corpus data and provides a single-page summary. The relations or grammatical relations are used to structure them in an orderly manner. *Figure 1* illustrates an instance of word sketch.

The system also offers the following features:

- More than 450 preloaded corpora for over 85 languages
- To streamline lexicography work; it offers tickbox lexicography application (Kilgarriff et al. 2010)

As indicated by Rychlý et al. 2008, algorithms are provided to suggest appropriate examples from dictionary

- The data from word sketch provides visual variances among words
- Word sketch data facilitates in deriving statistical thesaurus
- It creates frequency distributions in accordance with the provided criteria
- It defines sub-corpora and how to work with them
- It has a compilation of n-gram and word list.

The Corpus Architect incorporates corpus building facilities including:

- To tokenize and tag user corpora morphologically, for this purpose : it has built-in third party tools
- It provides different support functions to manage and share corpora
- It enables the users the build corpora from the web using URLs or keywords
- It builds corpora in different formats using users' data.

fire (noun) Alternative PoS: verb (3,176)		British National Corpus (BNC) freq = 14,172 (126.20 per million)			
modifiers of "fire"	3,292	0.90	nouns and verbs modified by "fire"	3,168	0.90
log	96	9.64	brigade +	282	11.11
a log fire			the fire brigade		
gas +	150	9.48	engine +	174	9.67
the gas fire			fire engine		
electric +	110	9.26	extinguisher	69	9.43
electric fire			fire extinguishers		
coal	71	8.65	escape	72	9.29
a coal fire			the fire escape		
open +	144	8.56	crew	78	9.01
an open fire			fire crews		
artillery	35	8.26	fighter	61	8.88
artillery fire			fire fighters		
accident	39	8.11	alarm	47	8.57
dobson v. general accident fire and life assurance			fire alarm		
forest	46	8.02	station +	167	8.37
a forest fire			the fire station		
barn	26	7.78	hazard	37	8.22
died in a barn fire			a fire hazard		
sniper	23	7.78	exit	25	7.82
sniper fire			the fire exit		
			verbs with "fire" as object	3,334	2.40
			light +	208	10.34
			catch +	197	9.34
			caught fire		
			open +	255	9.20
			opened fire on		
			set +	280	8.87
			set fire to		
			machine-gun	37	8.47
			of machine-gun fire		
			start +	124	8.33
			blaze	37	8.31
			a blazing fire		
			roar	29	8.02
			a roaring fire		
			extinguish	27	7.97
			to extinguish the fire		
			burn	37	7.83
			fire burned		
			fight	44	7.76

Figure 1: The word "Fire" - An instance of Word Sketch, Source: Sketch Engine

In the beginning, only monolingual corpora was decided to be supported by the system. However, as time passed by, the need for new features arose. Thus, the features accommodating multilingual and bilingual corpus data were introduced to Sketch Engine. Moreover, in the recent years it has become easier to access gigantic volumes of multilingual data, for instance with the help of public data sets of EUR-LEX, OPUS, and Europarl, due to which the bilingual and multilingual features of Sketch Engine appear apropos. There has been substantial improvement in the multilingual and bilingual features recently.

Manuscript published on 30 September 2019.

*Correspondence Author(s)

Sharipova Aziza Abdumanapovna, Tashkent University of Information Technologies named after Mukhammad al-Khwarizmi, Foreign Languages Department, Yunusabad, Badamzar str. 8-51, Uzbekistan, Yunusabad, Amir Timur str. 108.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>



The system was introduced with new features and some of the old features were significantly improved. In this paper, the significance of Sketch Engine with respect to bilingual functions is discussed. Furthermore, the usefulness of these functions, in the course of developing multilingual or bilingual studies and bilingual learners' dictionaries, is demonstrated.

II. MULTIPLE TYPES OF CORPORA

This paper uses different Sketch Engine terms which are explained as follows:

A. Comparable Corpora

This type of corpora are unaligned to each other. The same text translation is not included in them. Nevertheless, the types of text and its size is comparable, for instance a 50 million Uzbek newspapers corpus can be compared to 100 million American newspaper corpus, despite the fact that the news topics widely vary.

B. Parallel Corpora (Parallel Corpus)

It is multilingual or bilingual corpora that aligns monolingual sections to one another through paragraphs or sentences. Essentially, all the sentences in language 1 corresponding to all the sentences in language 2 are included in the data. It is possible that there may be segments that are untranslated and alignment is not 1:1.

C. Source and Target Language

As far as we consider bilingual learners' dictionaries, the meanings of source language and target language significantly vary. When a query is formulated in the start it is referred to as the source language, on the other hand, the results attained in a semi-automatic manner from other language are alluded to as the target language.

III. BILINGUAL CONCORDANCE

If a user wants to search in parallel corpora's aligned segments in accordance with different settings and wants the results to be displayed side by side, they can use a bilingual concordance.

1. Corpus Query Language includes Bilingual Queries

There is a pair of at least 2 monolingual corpora that are linked through meta-data in a parallel corpus, in Sketch Engine. An operator to bind the parts (two or more) within the operator are contained in the CQL (Corpus Query Language) that Sketch Engine uses (Jakubíček 2010). The syntax is as follows:

```
<query1> within <corpus2>:<query2>
```

Where;

The arbitrary CQL queries are represented by <query1> and <query2>, and the target language's aligned corpus is signified by <corpus2>. For instance, the query for Europarl Corpus's German part is as follows:

```
[lemma="Katze"] within europarl7_en: [lemma="cat"]
```

In the German part, all the occurrences of "Katze" are returned to the corresponding aligned segment containing "cat". This implies that cat is the translated result of "Katze", in most cases. The operator !within ('not within') can be used to search for the complement:

```
[lemma="Katze"] !within europarl7_en: [lemma="cat"]
```

The instances where "cat" is not the translated answer of "Katze" will be highlighted through this query, in technical words, such occurrences of "Katze" will be discovered when "cat" is not contained within the corresponding aligned section. These operators can be used to link three or more corpora.

IV. STATISTICAL TRANSLATION DICTIONARIES

For every pair of word in the corpus, with the help of sentence alignment; the users are permitted to measure the likelihood that word "A" will be translated to word "B", and various statistical computation over the aligned pairs is also possible. As indicated by Och and Ney (2003), when the training of translation model is done in SMT (Statistical Machine Translation), the procedure is quite the same as the one explained above. In light of the statement made by Rychlý (2008), the logDice implementation score (the same measure that word sketch uses in scoring collocation strength) is used in our implementation.

However, it depends on:

- The two words' standalone frequencies; the resulting score will be low if these frequencies are high.
- The two words' co-occurrence frequency (e.g. Cat and Katze); the resulting score will be high if this frequency is high.

The strongest "translation candidates" (in accordance with the scores) are listed with the help of the scores computed for pair of words. For each word, the top 10 candidates are stored as a statistical translation dictionary.

Sketch Engine utilizes MIS (Manatee Indexing System) that makes this process slightly less computationally challenging by exploiting an algorithm to compute bi-grams so that it is made easier for large corpora.

1. Translation Candidates are Highlighted

The translation candidates in the parallel concordance are highlighted through the statistical translation dictionaries. This is true even in the case when a query is not given, as Figure 2 illustrates. A specific target word is highlighted if source word's translation is found in the target section. The source word is also referred to as lemma. The dissimilarities between different pair of languages are massive, due to which the outcome is 85% successful on average, however the procedure that involves checking the concordance results is streamlined.

This feature is useful if the user wants to find out what a certain word may be translated as.

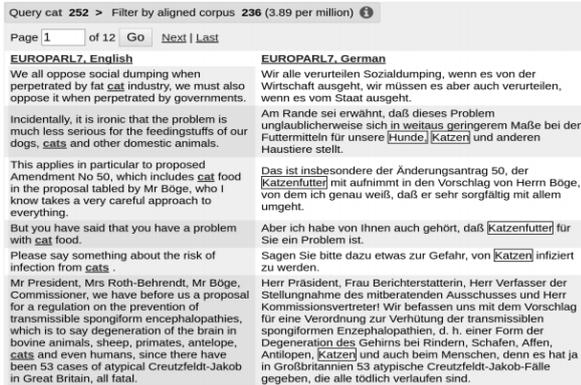


Figure 2: Statistical Translation Dictionaries

V. BILINGUAL WORD SKETCH

Sketch Engine assumes a noteworthy role when it comes to bilingual corpora. According to the founder of Sketch Engine, the working of bilingual word sketches is complex and it needs to be experimented with. In our case, these experiments have given rise to various variants of bilingual word sketch (Kilgarriff 2013).

This section describes the working and importance of these variants. As opposed to parallel concordances, both the comparable corpora and parallel corpora are facilitated by the bilingual words sketches' features.

A. Aligned Relations in Bilingual Word Sketch

In the past, to analyze the shared associations among words, two browser windows with word sketches for two separate languages, with two words providing translation of each other, were used. However, now the users are able to do this within a single window with the help of the basic version of bilingual word sketch. For the words in the target language and target corpus, there are new input boxes. The users are not permitted to select random target corpus in a random language. A combined word sketch, distinguished by color, for 2 words of 2 separate languages is shown as the result.

In older materials of Sketch Engine, the term BIM "Bilingual Manual" is used such a type of bilingual word sketch is referred to as BIM "Bilingual Manual" word sketch.

B. Translate Button

The dictionary already has ten good candidate translations stored thus the target language does not need to be inputted with translation by the users. This is made possible through the parallel corpus feature "automatic statistical dictionary". The most appropriate corpus for each target language may also be set in advance. The appropriate comparable corpus can be selected for non-parallel corpora, and the parallel section in the specific language can be selected for parallel corpora.

In this manner, the users are led to a "translate button" which is bilingual word sketch's further refined variation. The users are offered the translation of the word sketch into any available language as soon as a simple monolingual word sketch is created by them. The left panel in Figure 3 demonstrates the highlighted menu.

The comparable corpus and pre-configured parallel corpus are used by the function in statistical dictionary to find the most suitable translation and thus this word's bilingual word sketch is shown. If a user wants to translate in a different language or the selected choice is incorrect, they are offered links for other translations stored in the system's dictionary for the bilingual word sketch.

C. Compatible Relations

The compatibility of two specific relations is hard to determine. It should not be expected that one relation's collocation will be translated only to a particular relation's collocation in the target language, this is because in different languages, different things are implied through different means. Contrary to this, in the target language; one corresponding relation is used to translate majority of the collocations from a specific relation.

It is suggested that:

- The relations will be aligned to each other if they have identical names. For instance, in almost all the languages, the meaning of "modifier" will not greatly vary.
- UNIMAP is a directive in the word sketch grammar which can be used to map relations of the collocation to their nearest equivalents in English. This directive marks the compatibility of some relations with some relations in English. The two languages compatible relations (other than English) are the ones that show compatibility with the same English relation.

It is still quite complicated. An investment should be made for the advancement of compatible sketch grammars containing higher number of compatible relations. This will improve the overall corpora sketch application through the initial enhancement of the alignment.

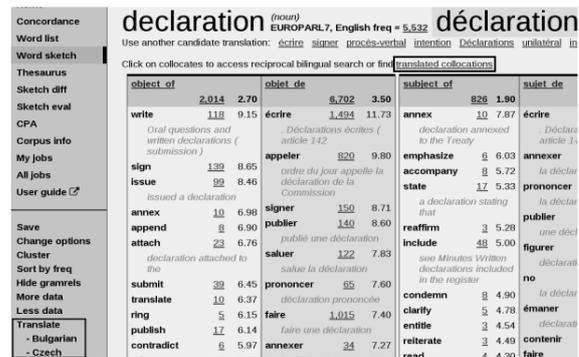


Figure 3: Bilingual Word Sketch

D. Bilingual Word Sketch - Reciprocal

It is highly likely that target language translates certain collocations while others remain untranslated when aligned relations are used to view a bilingual word sketch. Some of the collocations are not translated because in the target language, it is expressed using some other word. In such a situation, a direct way to access the word sketch of the specific problematic collocate is required by the users.

For instance, when a word sketch for Portuguese and English is viewed for *marrom* and *brown*, on the English side in the “modifies” relation; there will probably be *rice*, and on the Portuguese side; there will not be an equivalent. The users can click on *rice* using a small gadget in the interface, where *arroz*: a Portuguese translation can be inputted to get to the bilingual word sketch of *arroz* and *rice*, it will be revealed through this process that *brown rice* is an equivalent of *arroz integral* (i.e. no direct equivalent of *brown* is used at all). *Figure 4* demonstrates this function.

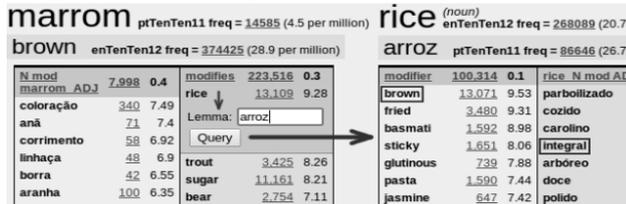


Figure 4: Bilingual Word Sketch – Reciprocal

E. Bilingual Word Sketch with Aligned Collocations

The collocations, in addition to headword, can also be translated through the automatic statistical dictionary. The automatic analysis is shifted slightly further by doing so; the system automatically searches for the translations of the target languages (in simpler cases), which saves the user from doing so manually. It is not necessary that the right translation exists within the same relation, because of this collocates are aligned across grammatical relations. In case of both target and source languages, for each of the collocations; examples are provided. The comparable corpora and parallel corpora can both use this feature. The only examples that are displayed for parallel corpora are the ones in alignment with each other, even if the data is presented by other collocations, they are not shown if they have no such example. *Figure 3* demonstrates the link of “translated collocations” under which the feature is available. *Figure 5* exhibits the translated collocations along with their resultant bilingual word sketch.

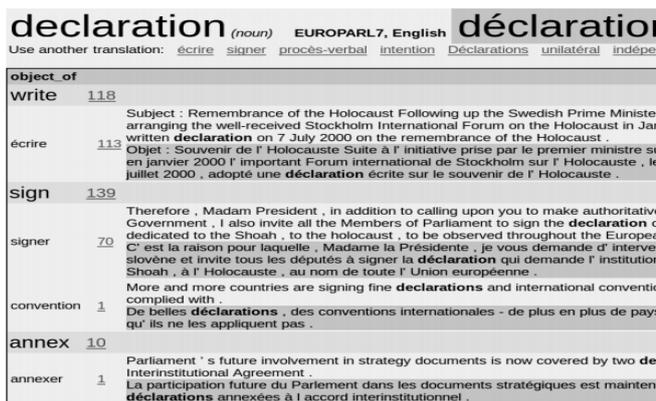


Figure 5: Aligned Collocations

VI. BILINGUAL TERMINOLOGY EXTRACTION

The automatic extraction of translation candidates by Sketch Engine for terms from specialized parallel corpus is described in this section. If translation candidates require to be drafted, then the system will do it automatically. Furthermore, if an entry list needs to be compiled for a specialized bilingual dictionary, then Sketch Engine can be used.

A. Terminology from a Monolingual Corpus

According to Kilgarriff et al. 2014, a specialized corpus contain term candidates which can be automatically extracted using Sketch Engine. To be specific, the appearance of the term is described through the queries in corpus query language (using morphological tags and specific words). The score is used for additional evaluation of candidates obtained through this procedure. To compute the score, the general corpus is compared against the uniqueness of the word for a specific type of text (Kilgarriff 2009). Following is the formula for score:

$$\text{Score} = (\text{freq_per_million} + 1) / (\text{ref_freq_per_million} + 1)$$

Where;

- In case of general corpus, the given language’s same frequency is represented by *ref_freq_per_million*.
- In case of specialized corpus, candidate’s per million frequency is represented by *freq_per_million*.

In a specific domain, the higher level of term’s specialness contributes towards the higher score on the basis of which candidates are arranged.

B. Alignment of Terms

As explained in Section 4, statistical dictionary facilitates in drafting process. However, it is possible to extend this process, so that instead of working with words, it is able to work with term candidates. For the alignment of terms, the extracted term candidates from every sentence is considered, and the co-occurrence score for T1 and T2 (term candidate’s pair) is computed in the same way as statistical translation dictionaries.

The source language contains term score which is used to sort the results. Moreover, the association score can also be used for this purpose. The particular domain’s specific terms are displayed at the top when term score is used, and in case of association score; the candidates with most salient translation appear at the top. *Figure 8* exhibits a sample output.

L1 term	L2 term	Co-freq	L1 freq	L2 freq
mobility	mobilité	23	27	25
retention	rétenion	49	60	51
consent	consentement	49	58	53
analysis	analyse	410	467	464
secondary school age	pourcentage du nombre	11	12	13
climate	changement climatique	102	122	110
hydration	hydratation	29	36	30
chain	chaîne	29	34	32
methodology	méthodologie	49	58	54
non-discrimination	non-discrimination	14	16	16
young woman	jeune femme	7	8	8
nutrition	nutrition	263	307	295
litre	litre	31	35	36
solid waste	déchets solides	24	25	30

Figure 6: The feature of bilingual terminology extraction

VII. M:N ALIGNMENT

Lastly, the 1:1 alignment of parallel corpora is not required by Sketch Engine anymore. Thus, 3:1 translation of 3 sentences into 1 is possible, and the system will not require alignment meta-data of 1:1. This plays an important role where granular alignment between two pairs is different, i.e. in case of multilingual parallel data. Moreover, the segments that are not translated can also be included, i.e. any of M, N can be 0. Nevertheless, it is still hard to set up such a parallel corpus as users are required to map file through sentence identifiers and finally linking this file in the configuration (Benko, V. 2014). Thus, even though users can utilize this machinery, it is most likely to be used for corpora that is loaded in advance in Sketch Engine.

VIII. CONCLUSION

Sketch Engine and its features relating to the corpus query system possess the capability of processing various languages, however it was mainly used as only a monolingual tool in the past. This is not the case anymore. There have been significant advancements related to working with bilingual corpora and bilingual lexicography in Sketch Engine. This paper provided a description of the role of Sketch Engine's features and principles. For the simplification of bilingual learner's dictionaries for the lexicographers, this research provided an overview of the features with examples, consequently contributing to the quality of their work.

REFERENCES

1. Baisa, V., Cukr, M. and Ulipová, B. 2015. Bilingual terminology extraction in Sketch Engine. In 9th Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2015), 65-70.
2. Baisa V., Michelfeit J., Medved' M. and Jakubiček M. 2016. European Union Language Resources in Sketch Engine. In Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016).
3. Benko, V. 2014. Compatible Sketch Grammars for Comparable Corpora. In Proceedings of the XVI EURALEX International Congress. Bolzano: EURAC research, 417-430.
4. Jakubiček, M., Kilgarriff, A., McCarthy, D. and Rychlý, P. 2010. Fast Syntactic Searching in Very Large Corpora for Many Languages. In PACLIC, 24, p741-747.
5. Kilgarriff, A. 2009. Simple Maths for Keywords. In Proceedings of Corpus Linguistics Conference CL2009. University of Liverpool.
6. Kilgarriff, A. Kovář, V. and Rychlý, P. 2010. Tickbox Lexicography. In eLexicography in the 21st century: New challenges, new applications. Brussels: Presses universitaires de Louvain, 411-418.
7. Kilgarriff, A., Kovář, V. and Frankenberg-Garcia, A. 2013. Bilingual word sketches: three flavours. In Electronic Lexicography in the 21st Century: Thinking outside The Paper (eLex 2013), 17-19.
8. Kilgarriff, A., Jakubiček, M., Kovář, V., Rychlý, P. and Suchomel., V. 2014. Finding Terms in Corpora for Many Languages with the Sketch Engine. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: The Association for Computational Linguistics, 53-56
9. Kilgarriff, A., Baisa V., Bušta J., Jakubiček M., Kovář V., Michelfeit J., Rychlý P. and Suchomel V. 2014. The Sketch Engine: ten years on. Lexicography, 1.1: 7-36.
10. Och, F.J. and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29.1: 19-51.
11. Rychlý, P. 2007. Manatee/bonito-a modular corpus manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007), 65-70.
12. Rychlý, P. 2008. A Lexicographer-Friendly Association Score. In 2nd Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2008). Brno: Masarykova Univerzita, 6-9.

13. Rychlý, P., Husák, M., Kilgarriff, A., Rundell, M. and McAdam, K. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Proceedings of the XIII EURALEX International Congress. Barcelona: Institut Universitari de Lingüística Aplicada, 425-432.
14. Abdumanapovna, S.A. 2018 The contemporary language studies with corpus linguistics, ACM International Conference Proceeding Series, 82-86 .

AUTHORS PROFILE



Doctor of Philosophy (PhD) and an Associate Professor of Science from 2019. He has published 42 works, including 1 monograph, 31 scientific papers (2 in scientific journals of Scopus database) and 9 manuals and a dictionary. And also she is the author of textbooks and mobile apps of the dictionary.

The mobile application of "Hardware in computing" tutorial on the practical application of the scientific work was positively evaluated and it was awarded "The Best Teacher of Higher Education Institution" Republican contest "The Best Teacher in the Field of Communication, Information and Telecommunication Technologies" in 2017.

She was the head of the English language course at the State Unitary Enterprise "Center for Radio Communication, Broadcasting and Television". In 2016-2017, she participated in the VII, an innovative project funded by the state budget at the Tashkent State Technical University named after Islam Karimov, to create the "Explanatory dictionary of terms and concepts in information security in Russian, Uzbek and English". She has also been a reviewer for dissertations, masters and graduate students in English.

A. Sharipova was the Organizer of the Republican Scientific-Practical Conference "Integration of Information Technologies and Foreign Languages in Globalization" which was held on February 26-27, 2019.

A. Sharipova is a researcher at other universities of the Republic (Urgench State University, Karakalpak State University named after Berdakh, Tashkent State Technical University, Uzbek State World Languages University) and academic lyceums (academic lyceum at Turin Polytechnic University in Tashkent, Academic Lyceum # 2 at Tashkent University of Information Technologies). Participates in the open lessons, seminars, round tables and shares knowledge and skills.