Dominic Stewart

A GUIDE TO SKETCH ENGINE FOR ITALIAN LEARNERS OF ENGLISH

TANGRAM
EDIZIONI SCIENTIFICHE
TRENTO

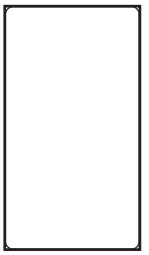
Dominic Stewart
A Guide to Sketch Engine for Italian Learners of English

Gruppo Editoriale Tangram Srl
via dei Casai, 6 – 38123 Trento
www.edizioni-tangram.it – info@edizioni-tangram.it

Series "Didattica" – SIN 22
First Edition: May 2024, *Printed in EU*

ISBN 978-88-6458-253-5

The volume was published with the contribution
of the Humanities Department of the University of Trento



Printing on ecological paper from forest areas, totally free of chlorine.
It does not contain optical brighteners, it is acid free with alkaline reserve

Contents

A GUIDE TO SKETCH ENGINE FOR ITALIAN LEARNERS OF ENGLISH

Introduction

What is Sketch Engine?

Sketch Engine[1] is a corpus manager and text analysis software developed in 2003. It allows people studying language behavior, e.g., lexicographers, translators and language learners, to search large text collections using linguistically motivated queries. For language learners it represents a very useful way of going beyond the perforce circumscribed information present in dictionaries.

At the time of writing Sketch Engine contains over 700 corpora in over 100 languages. Four main query types are available, namely Simple Query, Lemma Query, Phrase Query and Word Query, as well as Word Sketch, from which the name Sketch Engine derives.

Purpose of the book

The objective of this book is to explain in a clear and user-friendly way the main functions of Sketch Engine for learners of English studying at Italian universities and secondary schools, as well as for trainees of Italian-English translation. Having said that, it needs to be clarified from the outset that although (i) short excerpts of Italian tourist texts are adopted as a springboard for the various corpus investigations, and (ii) the book is sprinkled with comparisons of Italian / English cognates, this is not a translation textbook.[2] The choice of tourist texts as the object of study reflects a personal preference arising from many years of teaching the translation of the language of tourism, but many other non-technical registers would have been just as suitable.

[1] https://www.sketchengine.eu.

[2] If your principal requirement is a textbook of Italian to English translation, i.e., with texts to translate, comments and suggested solutions, then you could try Stewart, Dominic (2018), *Italian to English Translation with Sketch Engine: A Guide to the Translation of Tourist Texts*, Trento, Tangram.

Structure of the book

The chapters are divided as follows. Chapters 1 and 2 are prevalently quantitative in nature, since they are concerned with the calculation and the implications of frequency in English within the four main query types in Sketch Engine. Chapter 1 is devoted to frequency within the Simple Query, probably the most recurrently adopted query type, while Chapter 2 moves on to frequency within the other main query types, i.e., Lemma Query, Phrase Query and Word Query. The emphasis in Chapters 1 and 2 is on showing users how to select the right queries when investigating language use, because failure to make accurate distinctions between the query types usually leads to unhelpful results.

Chapter 3 focuses on some functions available to the user once the concordance is displayed, with particular attention to the Collocations function and the Filter function, and to the way in which these can furnish further insights into English usage.

In Chapter 4 the attention turns to Word Sketch, both monolingual (English alone) and bilingual (English / Italian cognates), with stress placed both on its advantages and disadvantages. The analysis will also include the option Sketch Difference, as well as multi-word queries in Word Sketch.

The conclusions will summarise the pros and cons of Sketch Engine for language learners.

Corpora adopted

Three corpora will be used throughout the book: the *British Web 2007* corpus, the *English Web 2021* corpus and the *Italian Web 2020* corpus. The *British Web 2007*, compiled in 2007, contains over 1.3 billion words derived from websites within the .uk domain. It is a general-purpose corpus with a broad range of text types. The more recent *English Web 2021*, compiled in 2021-2022, is much larger than the *British Web 2007* in that it has around 52 billion words. It too is a web-derived, general-purpose corpus with a broad selection of text types. The *Italian Web 2020* corpus has nearly 12.5 billion words, was compiled in 2019-2020, and again consists of texts collected from the

Internet. It will be adopted for the purposes of contrasting Italian usage with the results obtained from the two English corpora.

Why is it that the *British Web 2007* and the *English Web 2021* were preferred over the wealth of other English corpora available? The *British Web 2007* was chosen because it seemed important to have a strong native English element, notwithstanding the fact that .uk domain does not necessarily mean that the language users are all native speakers of English. The *English Web 2021* was chosen because in a sense it is in contraposition to the *British Web 2007*, i.e., it is very large, very recent and is more international, reflecting in greater measure the status of English as a global language.

For reasons of simplicity, during the book the three corpora will be referred to as the *British Web Corpus*, the *English Web Corpus* and the *Italian Web Corpus*, abbreviated in the tables to *BWC*, *EWC* and *IWC* respectively.

Dictionaries adopted

Two learner's dictionaries are adopted as reference works in this book: the online versions of the *Oxford Advanced Learner's Dictionary* and the *Longman Dictionary of Contemporary English*.

Proviso

It is worth stressing that writing a book about corpora and corpus software is a precarious business, because you are dealing with moving targets. New corpora are constantly compiled, and the software is constantly enhanced. A few years from now the interface and structure of Sketch Engine may look radically different, and of course this is all to the good. The result is that books describing a platform such as Sketch Engine can age quickly.[3] Nonetheless it is my hope that this work will prove beneficial to language learners for many years to come.

[3] For example, the guide to Sketch Engine produced by Thomas, James (2016), *Discovering English with Sketch Engine*, Versatile, was very useful at the time but is now harder to follow.

Procedural note

Note that throughout the book, italics are as a rule adopted for words, expressions and lemmas under analysis, but single inverted commas are adopted for corpus queries. This sometimes leads to an apparent repetition – for instance "to investigate the form *conditions* as a verb, use the Word Query 'conditions' and select Verb from the drop-down menu" – but in order to steer clear of misunderstandings it seems best to avoid overlap in this regard.

Chapter 1
Calculating frequency with Simple Query

1.0. Preliminary remarks

The first sections of this chapter examine the use of Simple Query to calculate the frequency of adjectives, nouns and verbs by comparing English / Italian cognates via the use of the English and Italian corpora specified in the Introduction. The rationale behind this is not only that contrastive frequencies can be more revealing than single frequency counts, but also that when we are confronted with words and expressions in an L2 it helps to be aware of their degree of frequency compared to equivalents or near-equivalents in our L1.

The subsequent sections of the chapter focus on the use of Simple Query to determine the frequency of orthographic variations, collocations, colligations and fixed expressions in English alone, and therefore only the two English corpora will be adopted.

Since the Simple Query is lemmatised, it retrieves all inflected forms of lemmas automatically, so for example the Simple Query 'easy' captures *easy, easier, easiest*, the Simple Query 'excursionist' captures *excursionist* and *excursionists*, and the Simple Query 'run' captures *run, runs, ran, running* (as well as *Easy, RUN* etc. because Simple Query is not case-sensitive and therefore captures both upper and lower case). Simple Query retrieves the different forms of a lemma within phrases too, so the query 'run the company' picks up all the forms of both *run* and *company*. Of course, some lemmas have just a single form, for example the adjectives *touristic* and *homonymous* analysed below, and therefore the respective Simple Queries 'touristic' and 'homonymous' (remember that single inverted commas are adopted for corpus que-

ries) retrieve only *touristic* and *homonymous*, given that English polysyllabic adjectives do not have inflected forms. Similarly, English adverbs are not normally inflected, though there are some exceptions (the comparative form *oftener* occurs 266 times in the *British Web Corpus* and over 8000 times in the *English Web Corpus*), and of course by definition fixed expressions such as *from time to time* do not inflect.

As outlined in the Introduction to this book, the focus in the current chapter and in Chapter 2 is almost exclusively on frequency outcomes. Consequently, you will find few explanations, or only partial explanations, to account for the frequency results identified, and this is because they are generally contingent upon issues and functions dealt with later in the book. Therefore, for the present a number of stones will be left unturned, but I shall return in subsequent chapters to many of the frequency findings in these first two chapters. We shall begin with adjectives.

As explained in the Introduction, initial excerpts will be provided in Italian as a stimulus for corpus searches.

1.1. Simple Query for frequency: adjectives

1.1.1. touristic vs turistico

Percorrendo la strada dopo altri due chilometri sulla destra, si giunge a un insieme di case disabitate [...] Questo risulta essere di interesse **turistico**, perché il paese, disabitato ormai da quarant'anni, offre ancora uno spaccato di vita [...]

You might wonder if it is possible to adopt the adjective *touristic* to render *turistico* in the above sentence. The first thing to note is that *touristic* is listed in native-speaker dictionaries of English, for example the online *Collins English Dictionary*, but not in the two British learner's dictionaries adopted in this book (*Oxford Advanced Learner's Dictionary* and *Longman Dictionary of Contemporary English*). The absence of this word from these two dictionaries might already put you on your guard, but you may be curious to know to what degree *touris-*

tic is adopted in English. The Simple Query 'touristic', which being a polysyllabic adjective retrieves *touristic* alone because it has no inflected forms, produces the following frequencies in the *British Web Corpus* and *English Web Corpus*:

touristic	Raw frequency	Frequency per million words
British Web Corpus	274	0.18
English Web Corpus	23,234	0.38

Since the *British Web Corpus* and *English Web Corpus* are of vastly different sizes – as noted in the Introduction the 2021 *English Web Corpus* is enormous – the raw frequency figures may not be immediately revealing. The frequency per million words, however, not only reduces the figures to something more manageable, but also enables a proportional frequency contrast between the data supplied for the two corpora.

The *British Web Corpus* result would suggest that *touristic* is neither especially rare nor especially frequent, but the total per million words in the *English Web Corpus* is twice as frequent as in the *British Web Corpus*, thus giving the impression that this adjective is adopted more habitually outside the UK. On the other hand, the figures for the Simple Query 'turistico' in the *Italian Web Corpus* are a lot higher. Bear in mind that since Simple Query is lemmatised, the query 'turistico' will capture not only the masculine singular *turistico* but also the masculine plural *turistici*, the feminine singular *turistica* and the feminine plural *turistiche*:

turistico	Raw frequency	Frequency per million words
Italian Web Corpus	1,121,743	77.28

The colossal difference in the frequencies of the two cognates across English and Italian is most readily accounted for by the fact that the English word *tourist* is massively used with adjectival function (*tourist industry, tourist attraction, tourist destination*), with the result that it occupies a lot of the potential terrain of the adjective *touristic*.

With regard to possible translations of *turistico* in the Italian passage above, since the text is intended for visitors to the places described, in reality it seems redundant to emphasise that the places are *di interesse turistico*. Once could begin the translation of the second sentence simply with *This is interesting because…*

1.1.2. homonymous vs omonimo

La cittadina di Resia si trova ai piedi dell'**omonimo** passo ed è circondata da magnifici paesaggi.

Again, the English adjective *homonymous* is present in native-speaker dictionaries but absent from our two British learner's dictionaries, and once again this provides an important clue as to its degree of frequency in modern British English. The Simple Query 'homonymous' generates the following frequency figures in the *British Web Corpus* and *English Web Corpus*:

homonymous	Raw frequency	Frequency per million words
British Web Corpus	46	0.03
English Web Corpus	5413	0.09

The figures are low, though as in the case of *touristic* above, it is noticeable that outside the UK domain the score is higher. Once again, note the greater frequency of the Simple Query score for the Italian cognate 'omonimo' (which again retrieves masculine and feminine forms in the singular and plural) in the *Italian Web Corpus*:

omonimo	Raw frequency	Frequency per million words
Italian Web Corpus	362,579	24.98

The above information, together with the fact that English *homonymous* is used primarily in specialised medical contexts (see section 3.1.4 below for further details) should persuade you to avoid adopting this adjective in non-specialist texts. To be preferred is *of the same*

name (2.32 and 3.47 per million respectively in the *British Web Corpus* and *English Web Corpus*). Thus the Italian excerpt above could be rendered with *The town of Resia lies at the foot of the pass of the same name*. Even simpler would be the repetition *The town of Resia lies at the foot of Resia Pass*.

1.2. Simple Query for frequency: nouns

1.2.1. excursionist vs escursionista

In lontananza appare il rifugio Podesteria, antica malga che racchiude una parte importante della storia della Lessinia: questa era infatti la residenza estiva del podestà di Verona. Il rifugio è un viavai di **escursionisti** […]

In the *British Web Corpus* and *English Web Corpus* the Simple Query 'excursionist', which captures both the singular and the plural forms, retrieves the following outcomes:

excursionist	Raw frequency	Frequency per million words
British Web Corpus	59	0.04
English Web Corpus	1847	0.03

It is clear that the English noun *excursionist*, singular or plural, is not widespread, and once again it is not listed in our learner's dictionaries. On the other hand, a search for the Italian cognate 'escursionista' achieves a much higher rate of frequency in the *Italian Web Corpus*:

escursionista	Raw frequency	Frequency per million words
Italian Web Corpus	49,616	3.42

Therefore, with reference to the (modern) Italian excerpt provided above it seems safer to render *escursionisti* with *hikers* or *hikers and cyclists*.

1.2.2. indication vs indicazione

Svoltando a destra si procede per 50 metri sulla SP66 fino a imboccare la strada con **indicazioni** per Sidolo e Ca' Scapini.

The two cognates *indication* and *indicazione* also show very different frequency outcomes across English and Italian, in that the Italian noun is four times more frequent than its English counterpart:

indication	Raw frequency	Frequency per million words
British Web Corpus	35,539	22.96
English Web Corpus	1,097,812	17.83

indicazione	Raw frequency	Frequency per million words
Italian Web Corpus	1,184,662	81.62

This must connect with the greater conceptual range of *indicazione*, which appears to cover much of the semantic area occupied by English nouns such as *information, suggestion, (road/map) directions, sign* and of course *indication*. In the Italian excerpt above, *indicazioni* corresponds to *directions* or *signs*.

Try making similar queries for *patrimony* and *patrimonio*.

1.3. Simple Query for frequency: verbs

1.3.1. explore vs esplorare

[…] una delle grotte più profonde al mondo con pozzi di centinaia di metri che raggiungono la profondità di quasi mille metri finora **esplorati**.

We now move on to verb forms, for which lemmatisation is particularly salient in view of the number of different inflected forms for each verb lemma, particularly in Italian.

Consider the respective frequencies resulting from the Simple Queries 'explore' and 'esplorare':

explore	Raw frequency	Frequency per million words
British Web Corpus	163,338	105.54
English Web Corpus	6,345,532	103.04

esplorare	Raw frequency	Frequency per million words
Italian Web Corpus	412,215	28.4

Here we encounter a reversal of the trend identified so far, because the English cognate is almost four times more frequent than its Italian counterpart in both the *British Web Corpus* and *English Web Corpus*. Can you hypothesise a reason for this?

The frequency rates of the two cognates appear to be contingent primarily upon their collocational ranges rather than upon the different meanings of the verb. The various inflected forms of the English verb seem to associate mostly with words such as *issue*, *way* (*of doing something*), *strategy*, whereas the multiple inflected forms of *esplorare* are prevalently associated with places. Having said that, since *explore* also combines regularly with places, it can assuredly be adopted to render *esplorati* in the passage above.

1.3.2. exaggerate vs esagerare

La strada che conduce alla baita è […] di una bellezza **esagerata**, avvolta nel sentiero e nei profumi degli abeti.

Take a look at the frequency figures of *exaggerate and esagerare* and try to account for the diversity across English and Italian:

exaggerate	Raw frequency	Frequency per million words
British Web Corpus	5,501	3.55
English Web Corpus	215,812	3.50

esagerare	Raw frequency	Frequency per million words
Italian Web Corpus	209,937	14.46

In this case there would appear to be a difference of conceptual range. While both verbs cover the area of the English verb *exaggerate*, the

Italian verb casts its net wider to include concepts covered by the English verbs *overdo, go too far, go overboard,* and is therefore much more frequent. For related comments on the adjectives *exaggerated* and *esagerato,* and therefore on a possible rendering of *esagerato* in the passage above, see section 4.1.3 under Word Sketch.

1.3.3. characterise vs caratterizzare

Il luogo è **caratterizzato** da un'antica fontana con quattro canali da cui scorre un'acqua limpida e fresca.

Once again you could check the frequency figures for the Simple Queries 'characterise' and 'caratterizzare', but unlike the above cases you need to be aware of the English spelling variants *characterise/characterize*. Fortunately in Simple Query it is not necessary to make two separate searches for these variants because you can simply write 'characteri?e', which will capture the two variants, again with all their inflected forms:

characteri?e	Raw frequency	Frequency per million words
British Web Corpus	31,553	20.39
English Web Corpus	1,447,402	23.50

caratterizzare	Raw frequency	Frequency per million words
Italian Web Corpus	1,627,731	112.14

There is no doubt that the Italian cognate is considerably more recurrent. Both the Italian and the English verb have a rather formal tone, something which you need to bear in mind when translating tourist texts from Italian to English.

Since English tourist texts are in general more informal than their Italian counterparts, you should be wary of overdoing *characterise / characterize* in your translations.

For an analogous search you could compare results for the queries 'valori?e' and 'valorizzare', which also reveal a significant frequency discrepancy.

1.4. Simple Query for frequency: orthographic variation

1.4.1. characterise vs characterize

As a development on the last query of the previous section, you might wish to know which of the two spelling variants – *characterise* or *characterize* plus their inflectional forms – is more frequent in the two English corpora:

characterise	Raw frequency	Frequency per million words
British Web Corpus	21,888	14.14
English Web Corpus	261,499	4.25

characterize	Raw frequency	Frequency per million words
British Web Corpus	9665	6.25
English Web Corpus	1,185,898	19.26

The figures clarify that the spelling *characterise* is a lot more widespread in the UK domain corpus, whereas in the more international *English Web Corpus* it is *characterize* which is by far the more recurrent. By way of comparison you could check out *analyse* and *analyze*.

1.4.2. medieval vs mediaeval

È possibile attraversare tramite due ponti, uno più moderno che permette il passaggio delle auto e uno **medievale** pedonale.

The Simple Query 'medi*val', where the asterisk stands for any number of unspecified characters, captures the two orthographic variants *medieval* and *mediaeval*. Which variant do you think is more recurrent across the two English corpora? Consider the outcomes for the Simple Queries 'medieval' and 'mediaeval':

medieval	Raw frequency	Frequency per million words
British Web Corpus	48,554	31.37
English Web Corpus	866,762	14.07

mediaeval	Raw frequency	Frequency per million words
British Web Corpus	4381	2.83
English Web Corpus	39.268	0.64

The first point to note is that both *medieval* and *mediaeval* are proportionally a lot more widespread in the *British Web Corpus* than the *English Web Corpus*, and the second is that *medieval* is much more prolific than *mediaeval* across the two corpora. It would seem, however, that either form could be adopted to render *medievale* in the Italian text above. Compare *encyclopedia* vs *encyclopaedia*.

1.4.3. plateaus vs plateaux; frescos vs frescoes

La strada della Forra, denominata oggi Provinciale 38, dalla Gardesana sale verso gli **altopiani** di Tremosine
Il suggestivo centro storico, ricco di costruzioni antiche e a**ffreschi**, sorge su di un'isola calcarea

You are now well-acquainted with the fact that Simple Query retrieves all inflected forms of a given word. Therefore, if for instance you make a Simple Query search for 'scholarship', both *scholarship* and *scholarships* will be returned. However, the two forms of *scholarship* are as a rule adopted in very distinct contexts. For this reason, if you are interested in the singular form alone it is more efficient to opt for the function Word Query and type in 'scholarship', because Word Query is enabled to capture single word forms (see section 2.3). If on the other hand you are interested only in the plural *scholarships*, it makes little difference which of the two queries – Simple or Word – you opt for, because in Simple Query the search for an inflected form (in this case 'scholarships') will return only the inflected form, and again Word Query captures the individual form, in this case the plural. This is contingent upon the fact that Simple Query, though lemmatised, is also able to analyse words that are not lemmas (the query 'scholarships' is a plural noun, not a lemma).

The above is relevant to the different plural forms *plateaus vs plateaux* (*altopiani* in the Italian text above) and *frescos vs frescoes* (*affreschi* above). Simple Query is adopted here because this is the query type under analysis in the current chapter, but Word Query would also be suitable. Let us begin with *plateaus* vs *plateaux*, i.e., two spelling variants of the plural form of *plateau*. Take a look at the outcomes of the Simple Queries 'plateaus' and 'plateaux':

plateaus	Raw frequency	Frequency per million words
British Web Corpus	333	0.22
English Web Corpus	28,105	0.46

plateaux	Raw frequency	Frequency per million words
British Web Corpus	430	0.28
English Web Corpus	5348	0.09

In the *British Web Corpus* the plural *plateaux* occurs slightly more often than the plural *plateaus* (if the record label Mille Plateaux is eliminated from the concordance then the frequency per million words is reduced from 0.28 to 0.25), whereas in the *English Web Corpus* it is *plateaus* that prevails. Now see if you can predict which is the more recurrent of *frescos* and *frescoes*. Below are the data for the Simple Queries of the two plural forms:

frescos	Raw frequency	Frequency per million words
British Web Corpus	164	0.11
English Web Corpus	10,384	0.17

frescoes	Raw frequency	Frequency per million words
British Web Corpus	879	0.57
English Web Corpus	47,152	0.77

Here it is the plural *frescoes* that occurs more prolifically in both corpora, though *frescos* is certainly attested. Incidentally, since Simply Query does not distinguish parts of speech (also known as 'word classes', but the information provided in Sketch Engine shows a terminological preference for 'parts of speech'), in these searches there was a

slight risk of retrieving *plateaus* and *frescoes* / *frescos* as present tense third-person singular verbs, though as it turned out, in the search results all the forms captured were nouns. See sections 2.1 and 2.3 on Lemma Query and Word Query for more information on querying parts of speech.

1.5. Simple Query for frequency: collocations

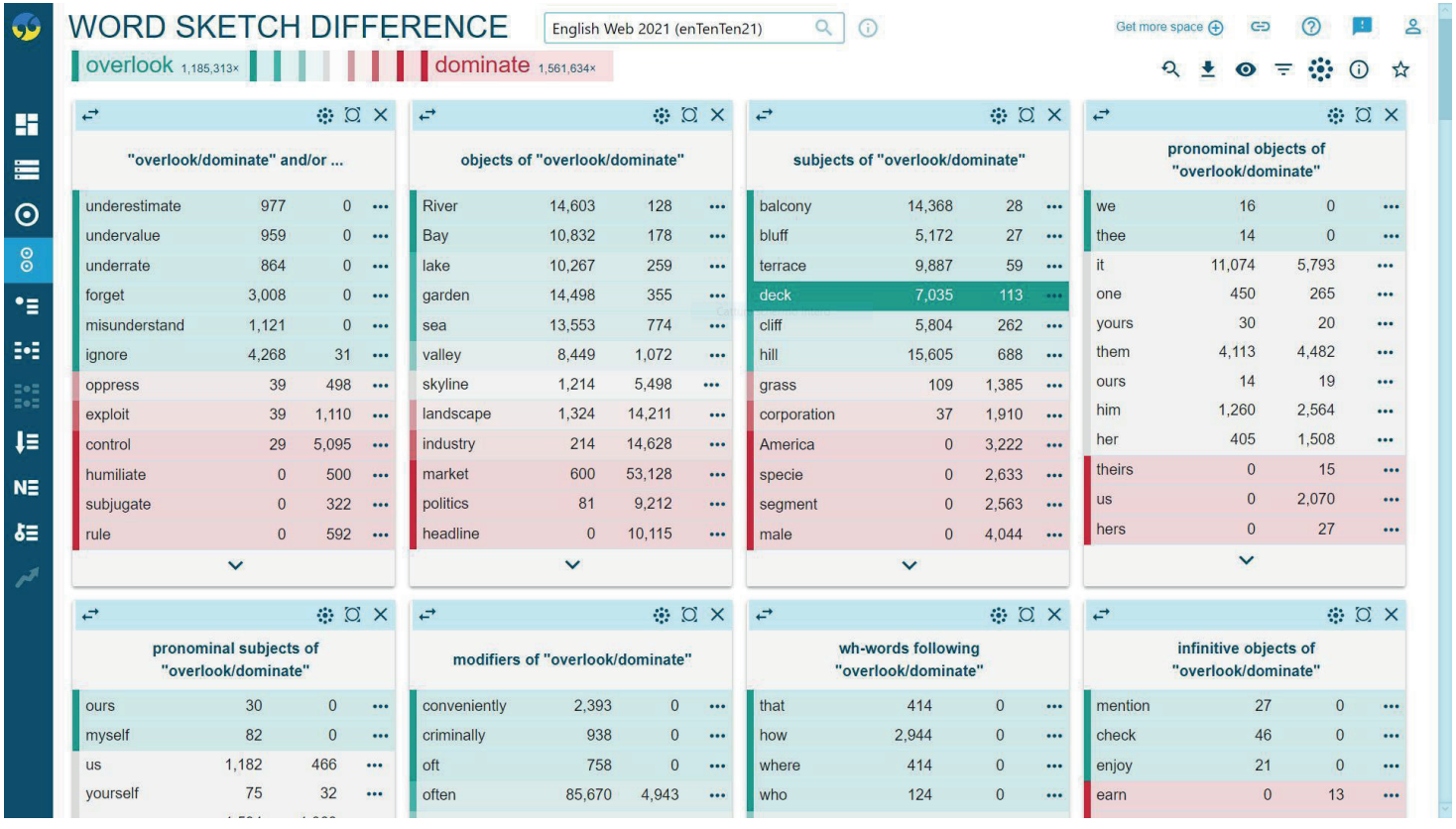
As observed in 1.0, Simple Query is enabled to retrieve not only single words but also sequences of words. In this section, two examples will be provided of how Simple Query can be utilised in order to investigate the frequency of collocations when it is the collocation itself that constitutes the query. This is to be distinguished from the Sketch Engine 'Collocations' function, which can be activated once a concordance has been displayed – see section 3.1 for details.

1.5.1. XVIII century vs 18th century vs eighteenth century

La chiesa di San Filippo, eretta **nel XVIII secolo** al di là del ponte, costituisce un monumento interessante dal punto di vista storico-artistico

Simple Query analyses word sequences in the exact order of the words typed. Therefore, the query 'furniture company' finds, for instance, *furniture company* and *furniture companies*, but not *company furniture*. Listed below are the frequencies per million words of various combinations involving the English word *century* in the *British Web Corpus* and *English Web Corpus*. Try to spot some patterns:

	British Web Corpus	*English Web Corpus*
xiv century	0.01	0.02
xv century	0.02	0.02
xviii century	0.03	0.05
xx century	0.03	0.09



Tab. 18 – Word Sketch Difference ‘overlook’ vs ‘dominate’ – EWC

Index of Words and Phrases Analysed