Iperstoria
2025, no. 25
ISSN: 2281-4582

DOI: 10.13136/2281-4582/2025.i25.1585
iperstoria.it

Lost in the Forest

Flashpoints for Language Learners in Sketch Engine

Dominic Stewart
University of Trento
ORCID: https://orcid.org/0000-0003-4447-0795
Email: dominic.stewart@unitn.it

Keywords

Word Sketch
Data-driven learning
Corpus queries
L2 acquisition
Lemma

Abstract

Recent work on data-driven learning manifests concern for the level of complexity that the use of corpora can entail, a complexity which – it is feared – may ultimately turn learners away from corpus interfaces in favour of the more discursive and user-friendly query strategies offered by AI. With this in mind, the primary aim of this paper is to trace the "journey of experience" of a hypothetical, relatively inexperienced, university-level user of the corpus management and analysis software Sketch Engine, in order to evaluate its plausibility as a resource for this type of user. Particular focus will be devoted to the Word Sketch option, following an analysis of the main differences between the query types Simple Query, Lemma Query, Phrase Query and Word Query. In this regard, it would seem that the issue of lemmatisation merits particular attention.

1. Introduction

1.1 Preliminary remarks

This analysis lies within the purview of data-driven learning (DDL), about which much has been written recently in the light of developments in the area of artificial intelligence. What is manifest in the literature on DDL is a preoccupation with the level of complexity that the use of corpora can entail, a complexity which – it is feared – may ultimately turn learners away from corpus interfaces in favour of the more discursive and user-friendly query strategies offered, for example, by ChatGPT.

The primary aim of this paper is to trace the "journey of experience" of a hypothetical university student using the corpus management and analysis software Sketch Engine, in order to reflect upon its plausibility as a resource for this type of user, with particular focus on the Word Sketch option. In this regard, it would seem that the issue of lemmatisation merits particular attention. The journey outlined is grounded in issues experienced by my Italian C1/C2-level university students of English when making basic queries.

I shall begin by examining the literature regarding the use of corpus interfaces on the part of learners, before moving on to Sketch Engine itself. Firstly I will discuss the notions of lemma and lemmatisation and how these apply to Sketch Engine searches; secondly I shall highlight the main differences between the query types Simple Query, Lemma Query, Phrase Query and Word Query; thirdly I will envisage the hypothetical user's experience within the Word Sketch function as they are confronted with some scarcely intelligible results stemming from what would appear to be rudimentary queries. Issues deriving from this experience will then be discussed.

1.2 Corpora in DDL

Corpora have a pivotal role in the teaching and learning of languages, either by means of hands-off methods, for example in the compilation of dictionaries or the creation of teaching materials, or by means of hands-on methods, whereby learners consult corpora themselves and are thus exposed to a greater extent to the language they are studying (Leech 1997). This latter method is known as DDL (Johns 1991; 1993), through which learners engage directly with corpus data, exploring and discovering patterns of usage (Crosthwaite and Baisa 2023; Gerigk 2023; Gilquin and Granger 2022). It is crucially intertwined with important pedagogical notions such as discovery learning (Flowerdew 2015; Bernardini 2004) and the cultivation of learner autonomy (Charles 2023; Gavioli 2009). As noted by Gilquin and Granger:

> By means of various activities […] learners are encouraged to observe corpus data, make hypotheses and formulate rules in order to gain insights into language (inductive approach), or check the validity of rules from their grammars or textbooks (deductive approach). They thus become more involved, more active and, ultimately, more autonomous in the learning process. (Gilquin and Granger 2022, 430)

See O'Keeffe (2021) for an overview of research on DDL.

Analyses of DDL have generally demonstrated its positive effects for language learning (Crosthwaite and Baisa 2023; Boulton and Cobb 2017), but a number of scholars report difficulties concerning the use of corpus tools and corpus data. It is claimed that a fair degree of technical expertise is often required for efficacious use of the available tools (Boulton and Vyatkina 2021; Crosthwaite and Cheung 2019, 171), inasmuch as good skills are necessary in order to manipulate the software and formulate appropriate queries (Crosthwaite and Baisa 2024; Leńko-Szymańska and Boulton 2015, 4; Boulton 2015, 270). The danger is that the absence of the necessary corpus literacy may result in users reaching flawed conclusions (Gilquin and Granger 2022). It is also important to note that the frequently complex interfaces may be at variance with how modern learners usually access digital information by means of

resources such as Google, ChatGPT etc. (Flowerdew 2024; Zadorozhnyy and Lai 2024; Boulton and Vyatkina 2021). The overall consequence of this – it has been claimed – is that corpus use may seem time-consuming and inefficient to less experienced users and thus generate frustration (Boulton 2024).

With this in mind, we will now turn to Sketch Engine, focusing above all on the formulation of queries and on the interpretation of the attendant outcomes.

1.3 What is Sketch Engine?

Sketch Engine[1] is a corpus manager and text analysis software developed in 2003. It enables searches of large text collections by means of linguistically motivated queries, and, for many users, it represents a useful way of going beyond the understandably limited information and examples supplied in dictionaries.

At the time of writing, Sketch Engine contains around 800 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, perhaps the jewel in the crown of Sketch Engine. The source of the examples provided in this paper is the recent, very large *English Web Corpus 2021*, which contains over 52 billion words. It is a general-purpose internet corpus with a broad range of text types.

2. The lemma in Sketch Engine

2.1 Defining a lemma

Sketch Engine queries are either lemmatised or non-lemmatised. This is a conventional dichotomy in corpus software, but it seems important to stress that lemmatisation/non-lemmatisation is the hub around which Sketch Engine queries revolve, and is therefore a crucial factor in determining successful interrogation of the corpora provided. However, the question is more complicated than it might seem to many corpus users.

The lemma is defined by Crystal as "an abstract representation, subsuming all the formal lexical variations which may apply: the verb *walk*, for example, subsumes *walking, walks* and *walked*" (Crystal 2008, 273). This description, though very concise, seems to reflect how the lemma is generally understood, i.e., like the headword of a dictionary entry, lemmas represent sets of word forms. This paper is not the place for an account of interpretations of the lemma in the literature, although the reader can consult, for instance, Knowles and Don (2004) for a discussion of some of those interpretations. What is crucial in the current context is how the lemma is defined in Sketch Engine, which is as follows:

[1] https://www.sketchengine.eu. Last visited 06/06/2025.

> The lemma is the form of the word found in dictionaries, sometimes called the base form. Introducing lemmas makes it possible to treat different word forms of the word as the same word. [...] The existence of the lemma makes it possible to type *go* and find *go, goes, going, gone* and *went* automatically. A wordlist generated on the lemma attribute will count the frequencies of *go, goes, going, gone* and *went* together and display them as one item: *go*.[2]

This definition seems analogous to the one suggested by Crystal, i.e., the lemma (*go*) as a representative of a set of word forms (*go, goes, going, gone* and *went*), although it gives less emphasis to the abstract nature of the lemma. For experienced language operators, the Sketch Engine definition is probably clear, but it seems legitimate to hypothesise that the less expert user would struggle with (i) the way the definition is formulated and (ii) the use of the term "base form."

Regarding (i) the formulation of the Sketch Engine definition, the first sentence refers to the lemma as the "base form" of the word. If we incorporate this information into the second sentence, replacing "lemmas" with "base forms of words," the outcome is as follows:

> Introducing base forms of words makes it possible to treat different word forms of the word as the same word. [my formulation]

This is even more repetitive than the original Sketch Engine definition, but it renders the ambivalence of the original more explicit, whereby lemmas and word forms appear to overlap. It goes without saying that this overlap is undesirable, because lemmas and word forms should be clearly distinguished: making a lemma query and making a word form query are two radically different procedures. For example, the Lemma Query 'go' retrieves very different outcomes by comparison with the (unlemmatised) Word Query 'go',[3] which captures only the single form *go*.

Moreover, (ii) the use of the term "base form" in the Sketch Engine definition above ("The lemma is the form of the word found in dictionaries, sometimes called the base form"), referred to as "basic form" elsewhere in Sketch Engine, may also puzzle the inexperienced user, because "base form" can also denote *solely* the uninflected form of a word, i.e., to the exclusion of any other forms. Moon (2010, 208-209), for example, reproduces a sample concordance of the lemma *know* in a spoken corpus, underlining that "94 per cent of occurrences are as the base form *know*" (2010, 209), and her examples include "we don't really know," "you know how to balance them,"

[2] https://www.sketchengine.eu/blog/words-tags-lemmas-lemposes-lowercase/. Last visited 06/06/2025.
[3] Throughout this paper, italics are adopted for words, expressions and lemmas under analysis, but single inverted commas are adopted for corpus queries. This can lead to an apparent repetition – for instance "to investigate the form *damages* as a noun, the user can make the Word Query 'damages' and select Noun from the drop-down menu" – but in order to steer clear of misunderstandings it seems best to make a clear graphic distinction.

"we were chasing you know maybe fighting the Belgians." Clearly, in these examples, Moon's reference is not to the lemma, but to the single uninflected form *know*.

The fact that "base form" can also correspond to "uninflected form" rather clouds the Sketch Engine definition of "lemma" reported above, because if "base form" is interpretable both as a lemma and as an uninflected form, then confusion may ensue. The Sketch Engine team informs the user that a lemmatised query searches for the base form (i.e., lemma), but if the user construes base form as corresponding to uninflected form, then this is precisely the type of search that lemmatised queries are *not* designed to perform. Both Simple Query and Lemma Query are lemmatised, and neither of them is enabled to retrieve, for example, the word form *explode* alone. The Simple Query 'explode' and the Lemma Query 'explode' automatically retrieve all forms of the verb, that is, *explode, explodes, exploded* and *exploding*.

Of course, it could be counter-argued that less experienced users do not need to reflect upon the nuances of base form, basic form and lemma; that, on the contrary, they simply need to grasp what a lemma *does* in Sketch Engine before undertaking searches. This argument may be justified with reference to simpler searches, but, as will be posited in the following sections, once the user scratches the surface of Sketch Engine queries, an understanding of the exact meaning of "lemma" is indispensable.

2.2 Which types of words qualify as lemmas in Sketch Engine?

It is evident that theorising the lemma would be redundant if all words had just a single form. Lemmas are not word forms, but their raison d'être is predicated upon the existence of word forms. When scholars define or discuss lemmas, there is understandably a tendency to exemplify them with those that represent multiple forms (by the standards of English), such as the irregular verbs *go* or *take*. Indeed, it seems almost counter-intuitive to associate lemmas with single-form words such as English prepositions and articles. As a result, it may escape the Sketch Engine user's attention that lemmas can belong to any word class, even to numerals, so the conjunction *but*, the pronoun *she* and the numeral *eighty-two* (but not *82*) are classified as lemmas in Sketch Engine, and are therefore acceptable as Lemma Queries. Indeed, it would seem that the only words that are rejected by Lemma Query are inflected word forms, for instance *wants* (whether as noun or verb), *faster* (adjective or adverb) and *scientists,* or the genitives *scientist's* and *scientists'*. Although this is perfectly rational and legitimate as a methodology, it is interesting to note in passing that those words for which the theorisation of the lemma is of no relevance or purpose (*but, from*) are eligible for lemmatised searches, whereas those words whose existence creates the need for theories of the lemma (*exploded, explodes,* etc.) lie outside the remit of lemmatised searches.

3. Sketch Engine queries

3.1 Simple Query and Lemma Query

As suggested by its name, Lemma Query investigates lemmas, but since lemmas can be typed in Simple Query too, the user might wonder why Lemma Query exists at all. Indeed, in terms of lemmatised searches, Simple Query is, in a sense, the more powerful of the two, because it accepts multi-word queries. If students of English wish to know whether roads can be said to climb, for instance, "the road climbed steadily up the hill" [my example], they can type the Simple Query 'road climb', which in the *English Web Corpus 2021* captures *road climb, road climbs, road climbed, road climbing, roads climb, roads climbed* etc. (but not *road was climbing*, *road then climbs* etc.) Queries comprising longer sequences of three or more lemmas such as 'road climb steadily' are also admissible in Simple Query.

By contrast, since Lemma Query is enabled solely for searches consisting of one word, 'road climb' and 'road climb steadily' are not admissible, a detail which might wrong-foot the less experienced user inasmuch as *road, climb* and *steadily* are all lemmas. As a result, if Simple Query offers a broader spectrum of lemmatised searches, what is the purpose of having another lemmatised query type?

A first answer is that Simple Query accommodates both lemmas and word forms; therefore, the queries 'climb', 'road', 'climbs' and 'roads' are all legitimate. However, it is worth stressing that if the user types 'climb' or 'road' in Simple Query, each of which is in theory interpretable both as a lemma and as a word form, the software will automatically construe them as lemmas, capturing – as nouns or verbs – *climb, climbs, climbed, climbing*; *road, roads, roading, roaded* etc. In other words, the search for the lemma overrides the search for the word form, a further detail that would merit more emphasis within the Sketch Engine website description of Simple Query, because it may be far from obvious to less expert users. By contrast, in Lemma Query, only lemmas can be typed, so 'climbs' and 'roads' are not possible.

A second answer is that, unlike Simple Query, Lemma Query offers the part of speech option. Therefore, if users wish to know what type of things climb, in Lemma Query they can begin by making the query 'climb' as a verb, a strategy which will exclude the prolific number of occurrences of *climb* as a noun in the corpus. Simple Query cannot handle part of speech queries.

Therefore, Lemma Query is rather more specific and focused than Simple Query.

3.2 Phrase Query and Word Query

The other two main query types – Phrase Query and Word Query – (for Word Sketch, see below) are not lemmatised. They simply return exactly what is typed in the query, so 'damage' retrieves only the word form *damage*, and 'damages' retrieves only the word form *damages*. There are, however, two fundamental differences between them. Firstly, Word Query offers a drop-down

menu for part of speech, whereas Phrase Query does not, so Word Query is enabled to retrieve the form *damage* only as a verb or only as a noun, or the form *damages* only as a verb or only as a noun. Secondly, Phrase Query can handle both one-word and multi-word searches, e.g., 'damages', 'considerable damages', 'very considerable damages' etc., while in Word Query it is possible to type only single words.

4. Summary so far

Some of the elements discussed above may catch less experienced users of Sketch Engine unawares. While it is true that the Sketch Engine team provides a range of descriptions, instructions and screencasts to assist users, the explanations regarding the differences between the four query types discussed above are somewhat concise. This is no doubt for cogent reasons, since elaborate descriptions may disconcert users, but the consequence is that a number of factors remain unemphasised. In the preceding sections it has been underlined that in Sketch Engine:

- defining a lemma as a "base form" may be misleading
- the Sketch Engine definition of "lemma" could be formulated more clearly
- any part of speech is classifiable as a lemma, including conjunctions and numerals

To this, it may be added that the two query types enabled for multi-word searches (Simple Query and Phrase Query) preserve the word order of the search in the results, in the manner of a Google query placed within inverted commas, e.g., the Simple or Phrase Query 'really actually' captures *really actually* but not *actually really*.

The differences between the four query types may be schematised as follows:

	Lemmatised?	Part of speech?	Multi-word option?
Simple Query	✓	*x*	✓
Lemma Query	✓	✓	*x*
Phrase Query	*x*	*x*	✓
Word Query	*x*	✓	*x*

Tab. 1: Differences between the four query types

If we recall another fundamental divergence, i.e., that Lemma, Phrase and Word Query are case-sensitive whereas Simple Query is not, it seems unlikely that even a more proficient user would have total awareness of all the distinctions outlined above without a degree of reflection. Therefore, it seems fair to assume that those distinctions might well escape users with less expertise in the use of corpus interfaces.

5. Word Sketch

5.1 What is Word Sketch?

The Sketch Engine website furnishes the following description of Word Sketch:

> The word sketch processes the word's collocates and other words in its surroundings. It can be used as a one-page summary of the word's grammatical and collocational behavior. The results are organised into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb, words that modify the word etc.[4]

In short, Word Sketch offers an overview of the environment of a word by adopting columns containing grammatical relations, within which frequent lexical and grammatical combinations are foregrounded.

5.2 A user's journey through Word Sketch

Let us now turn to the hypothetical university student's journey of experience mentioned in the Introduction. The scenario envisaged is that of a relatively inexperienced user of Sketch Engine and of corpora in general, trying to come to terms with Word Sketch in particular, although other query types will also be referred to.

5.2.1 Stage 1 of the journey: seeked vs sought

As observed in 3.1, Simple Query is able to handle searches featuring one or more lemmas (*fast deliveries, fast delivery*), as well as searches featuring one or more word forms (*faster delivery*, *faster deliveries*), whereas Lemma Query is enabled exclusively for queries consisting of one lemma, so word forms and multi-word queries are not admissible.

When users turn to the Word Sketch interface, they find – in faint type – the word *lemma* in the query space. This instruction could perhaps be delivered more forcefully, because the user guide to Word Sketch almost always refers to *words* that can be searched and not to lemmas – in the relevant screencast,[5] lemmas are not mentioned once – and the function itself is, after all, named Word Sketch rather than Lemma Sketch (see also the definition of Word Sketch in 5.1 above). This is the first of a series of potential pitfalls to be outlined within the journey described here.

[4] https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations. Last visited 06/06/2025

[5] "Word Sketch – analyse collocations in a corpus." https://www.youtube.com/watch?v=tKdEa1E-p2Y. Last visited 06/06/2025.

To begin with, imagine our user of Sketch Engine who has come across an example of a native speaker of English using the form *seeked* as the past tense of *seek*. Having previously been taught that the past tense (and past participle) form of *seek* is *sought*, the user is surprised by this discovery and wishes to ascertain how recurrent *seeked* is, and in which contexts and environments it is attested.

1. The user begins with the Simple Query 'seeked' and finds more occurrences (3236) in the *English Web Corpus 2021* than expected, deciding, as a result, to compare the uses of the verb forms *seeked* and *sought* in Word Sketch. Now, this is already a mistake, because only lemmas – and not word forms – can be entered in the Word Sketch query. However, it is perhaps an understandable mistake if we consider that (i) the definition of lemma in Sketch Engine is not crystal-clear (see 2.1 above), (ii) the descriptions of Word Sketch on the Sketch Engine website are generally framed in terms of words rather than lemmas, and (iii) even if the user has assimilated that Word Sketch queries allow only lemmas, they might assume that Word Sketch functions like another lemmatised query, such as Simple Query, which admits both lemmas and word forms.

2. Having typed in the respective Word Sketch queries 'seeked' and then 'sought' as verbs (the part of speech option is available for Word Sketch), our language learner discovers that the outcomes are 1720 occurrences of *seeked* as a verb and zero occurrences of *sought*. If the search were to stop here, then the conclusion might be reached – however perplexing it might seem – that *sought* is never used in English, and that *seeked* is not particularly frequent (remember that the *English Web Corpus 2021* is very large). This recalls the observation by Gilquin and Granger (2022), mentioned in section 1.2, about users reaching flawed conclusions.

3. However, curious to know the reason for the discrepancy in the number of hits for 'seeked' across Simple Query (3236 hits) and Word Sketch (1720 hits as a verb), the user clicks on the drop-down menu in the Word Sketch results column, just below 'seeked as verb', making the startling discovery that, aside from the 1720 occurrences as a verb, *seeked* is (automatically) tagged nearly 1500 times as an adjective, and that there is even a handful of occurrences of *seeked* as a noun. This means that Word Sketch has turned up almost as many occurrences of *seeked* as Simple Query, which is a decidedly mysterious outcome because, as stated above, Word Sketch is enabled to find lemmas alone, so, in theory, the results for *seeked* in Word Sketch should be zero.

4. Now unconvinced by the zero occurrences of *sought* in Word Sketch, the user returns to Simple Query and types in 'sought', which turns up nearly 2 million instances of *sought*. At this stage, one imagines a degree of head-scratching on the part of the user. The source of the problem is a misguided Word Sketch query – though, at the current stage of the journey, the user remains unaware of this – compounded by some apparently defective tagging.

The issues raised so far will be taken up again in Stage 5 of the journey (5.2.5) on Word Sketch Difference, but, for the present, we shall move on to the second stage of the user's journey of experience.

5.2.2 Stage 2 of the journey: the phrasal verb take after

The journey has already turned up some potentially mystifying outcomes. However, let us now suppose that the user, undeterred, is interested in the environment of the phrasal verb *take after*. Aware that Word Sketch, like Simple Query and Phrase Query, can accommodate multi-word searches, the user makes the Word Sketch query 'take after'. Further surprises are in store.

1. In the results, the word order of the query undergoes inversion, i.e., *after take*, of which there are over 3000 occurrences, the most recurrent sequences being (i) *after take off*, e.g., "we can have you at Battersea heliport just 15 minutes after take off," (ii) *soon after take*, e.g., "as the lads from Liverpool soon after took the stage," (iii) *shortly after take*, e.g., "his journey shortly after takes him to Russia," and (iv) *take after take*, e.g., "the director took take after take of the same scene." Strikingly, there is no trace of the phrasal verb *take after* among the results.

2. The user may need a while to absorb this. Casting around for explanations, their speculations might be as follows:

- Word Sketch accepts multi-word queries and distinguishes parts of speech, but it is not possible to specify the required parts of speech of each constituent of a multi-word query, so the user cannot specify that the *take* of the query 'take after' is a verb.
- Word Sketch accepts multi-word queries, but unlike other multi-word query types (Simple Query, Phrase Query, see section 4 above), its results can entail inversion, or at least a different order, of the words typed into the query. Here, the user's speculation is correct, because it is perfectly normal for a Word Sketch query to locate results with a different word order. For example, the Word Sketch query 'look up' finds not only the phrasal verb but also usage such as "the grassland higher up looked good" and "My only real complaint here is that the alien straight up looked like a rubber prop you would get at Halloween"; 'fill out' captures "a long hallway stretches out filled with glass cases" and "an alternative Halloween night out filled with high-energy live music"; 'look after' finds many occurrences of *after looking*; finally, 'badly park' locates, among others, both *badly parked* and *parked badly*.
- Word Sketch deals with lemmas, and the phrasal verb *take after* features two lemmas – *take* and *after* – so why doesn't the search 'take after' retrieve the phrasal verb?

3. Still casting around, the user makes the same query again, but this time clicks on the small question mark above the query space and identifies previously missed information: unlike both

Simple Query and Lemma Query, Word Sketch admits only lemmas which are nouns, verbs, adjectives and adverbs. It is worth stressing that this is an entirely new criterion, since no other Sketch Engine query makes this distinction.

4. At this point, the fog begins to lift, because the prioritisation of nouns, verbs, adjectives and adverbs would seem to account for the fact that *take after*, when it corresponds to verb + preposition, is not generated by the Word Sketch query 'take after'. However, if that is the case – the user reasons – why is *minutes after take off*, where *after* appears to be a preposition, retrieved 195 times by the same query?

5. One can imagine that the now rather bemused language learner indulges in some earnest head-shaking, because conclusive explanations for the outcomes retrieved so far remain elusive. However, it occurs to this most dogged of users to check the tagging of the 457 occurrences of *after take off* retrieved by the Word Sketch query in question ('take after'). Another surprise awaits: in the exact sequence *after take off*, not only is *take* always tagged as a verb, but *after* is always tagged as an adverb. It is for this reason that Word Sketch 'take after' retrieves *after take off*, i.e., because it is enabled for those lemmas which are tagged as nouns, verbs, adjectives and adverbs. *After take off* is retrieved because, in this sequence, *after* is tagged as an adverbial lemma.

6. At this juncture, in order to shed further light on the situation, the user who has not already given up may further check out the tags assigned to the sequence *after take off* by adopting Phrase Query, which is not lemmatised and retrieves exactly what is typed. The outcome of the Phrase Query 'after take off' is interesting: figures show that, of the 2214 instances retrieved, around 75% are tagged as 'preposition + verb + particle', whereas 25% are tagged as 'adverb + verb + particle', i.e., in 25% of the occurrences, *after* is labelled as an adverb.

Since Word Sketch cannot admit prepositions within queries, the Word Sketch query 'take after' does not capture the 75% of instances of *after take off* in the corpus where *after* is tagged as a preposition. This is a crucial point. In English, phrasal verbs consist of a verb and a particle, which may be a preposition ("she looked up a word in the thesaurus") or adverb ("at last things are beginning to look up"). However, the Word Sketch 'look up' captures only those instances where the tagger labels *up* as an adverb.

5.2.3 Stage 3 of the journey: after takeoff

Nonplussed by the apparently inconsistent tagging of the outcomes of the Phrase Query 'after take off', the user nonetheless recalls that the noun *takeoff* is usually written as a single word (*takeoff*) rather than two separate words (*take off*), and therefore decides to check out the sequence *after takeoff*.

1. Word Sketch dismisses the query 'after takeoff' with the message "Something went wrong… An error occurred while loading the data," which is disconcerting, because it gives the impression that there is some sort of temporary technical defect and that it is advisable to try again at some other time. This could result in repeated attempts over a number of days and a consequent waste of time, energy and patience. However, let us assume and hope that the user is still sufficiently *compos mentis* to posit that, in all probability, the message "Something went wrong…" appears because the sequence *after takeoff* involves a preposition (preposition + noun). The user therefore inserts 'after takeoff' in Simple Query, which generates 11,295 occurrences of *after takeoff*, and 5 of *after takeoffs*. By clicking on 'Frequency' above the concordance and then on the KWIC (Key Word In Context) 'Part of speech', the user establishes that all of the occurrences of *after takeoff* are indeed tagged as preposition + noun. This explains why the Word Sketch Query 'after takeoff' produces zero outcomes.
2. At last, there is a chink of light through the fog. With regard to the original Word Sketch query 'take after', the automatic tagger must have been deceived by the two-word spelling – within the sequence *after take off* – of the noun *takeoff*, as a result interpreting (but only 25% of the time) *take* as a verb and the preceding *after* as an adverb qualifying the verb.

Sadly, the discovery process has been laborious, frustrating and at times barely comprehensible, tantamount to a kind of survival of the fittest, or at least of the most tenacious.

5.2.4 Stage 4 of the journey: seek after

Perhaps as a kind of serendipitous side-search following the queries 'seeked' and 'take after', the user decides to compare the results in Word Sketch and Simple Query of the search 'seek after'. Word Sketch generates 1150 occurrences: *after* is always tagged as an adverb, and the most recurrent sequence in the columns is *highly sought after* (207 hits). However, the same search in Simple Query ('seek after') captures nearly 165,000 occurrences, and in just under 164,000 of these, *after* is labelled not as an adverb but as a preposition, with the combination *highly sought after* occurring around 28,000 times. Once again, there is the considerable risk of being misled by the Word Sketch outcomes, i.e., the user may conclude that 'seek after' occurs only around 1000 times in the corpus, whereas in in fact it occurs 165,000 times, and that *highly sought after* occurs only 207 times, whereas in fact there are around 28,000 occurrences. Further, in a collocation such as (*highly*) *sought after,* it will probably not be clear to users in the first place to which part of speech (adverb or preposition) *after* actually belongs, a situation which makes interpretation of the data even more arduous.

A final observation may be made about *after* in the *English Web Corpus 2021* in Word Sketch. The query 'after' with unspecified part of speech ('auto') is unsuccessful:

> Word Sketch cannot analyse this word. Some parts of speech cannot be analysed. Typically, only nouns, adjectives, verbs and adverbs are supported.

However, the same Word Sketch query 'after' with part of speech 'adverb' produces over 528,000 occurrences (8.59 per million) in the corpus. This specific discrepancy may be the result of a bug, but by this stage of the journey, it seems improbable that the user would be sufficiently lucid to distinguish bugs from all the other puzzling outcomes encountered.

5.2.5 Stage 5 of the journey: seeked vs sought in Word Sketch Difference

We can assume that, as a consequence of insights gained during the previous stages of the journey, the user has understood that Word Sketch queries reject word forms (e.g., *deliveries, faster*). However, the user may be taken aback upon discovering that, on the contrary, the Word Sketch Difference function includes them. As explained on the Sketch Engine website, Word Sketch Difference compares the collocates (i) of two different lemmas, for example 'damage' vs 'harm' (either as nouns or as verbs), or (ii) of two different word forms belonging to the same lemma, for example the word forms 'damage' vs 'damages' belonging to the noun lemma *damage*, and the word forms 'seek' vs 'seeks' belonging to the verb lemma *seek*. This second option allows the user to discover, for example, that (i) *damage* and *damages* as nouns have very distinct collocational environments, and (ii) the uninflected *seek* has a much stronger collocation with *advice* or *help* than *seeks* does, whereas *seeks* has a much stronger collocation with *attorney* or *applicant* than *seek* does.

The above searches work perfectly in Word Sketch Difference. Since the forms *seeked* and *sought* also belong to the lemma *seek*, one would expect this query to produce results as well, but more unexpected outcomes lie ahead.

As instructed, the user types in 'seek' on the line reserved for the lemma, then enters 'seeked' and 'sought' respectively in the two lines reserved for the 'first word form' and 'second word form', selecting 'verb' from the part of speech menu. However, this query generates the message "Nothing found. Please try to change search criteria," plus the additional indication "No results found for 'seek.'"

This baffling outcome is likely to persuade the user to give up altogether. If no results are found for the lemma 'seek' here, then why was the previous search (lemma 'seek': word forms 'seek' vs 'seeks') successful?

One envisages further speculation on the part of the frustrated user, but it is unlikely that they will find an answer. In fact, the Word Sketch Difference query is unproductive because the form *seeked* as verb – as well as *seeked* as an adjective (1482 hits) and as a noun (17 hits) – *is always classified as a lemma* in the corpus, and therefore the attempt to capture *seeked* only as

a word form is unsuccessful. This finding is so bizarre that it is worth stressing again: the word form *seeked* is always categorised as a lemma in the *English Web Corpus 2021*, despite the fact that, when tagged as a verb, it is labelled VVD (past tense) or VVN (past participle). One can only hope that this is a rogue bug, because it collides with the logic and criteria of lemmatisation.

6. Discussion

During the journey outlined above, the user's fairly basic Word Sketch queries in the *English Web Corpus 2021* have run aground because of:

- poor queries: 'seeked' and 'sought' (because Word Sketch rejects word forms); 'take after' to try to capture instances of the phrasal verb *take after* (because Word Sketch rejects prepositions); 'after takeoff' (again because Word Sketch rejects prepositions)
- poor spelling in original texts or in transcribed oral texts, which in turn generates unexpected tagging: 'after take off'
- inconsistent tagging: 'after take off' (*after* is sometimes tagged as an adverb and sometimes as a preposition)
- erroneous tagging: 'seeked' (always categorised as a lemma)
- bugs: 'after' as an adverb generates zero outcomes, even though Word Sketch is enabled for adverbs, as well as nouns, verbs and adjectives
- ambiguous messages: "Something went wrong… An error occurred while loading the data"
- the fact that, in certain cases, the user understandably lacks sufficient awareness of parts of speech: *after* in the sequence *highly sought after*

Notwithstanding the above, it could be argued that most of what has been presented so far in this paper is the result of a paradox: although the intricacies of Sketch Engine can certainly baffle language learners even at C1-C2 level, this resource is ultimately not designed for language learners. More specifically, the argument might go, Sketch Engine is a research tool rather than a learning/teaching tool. It supports lexicography, linguistics, collocation resources, translation, text prediction etc., but was not created with hands-on learners in mind. Indeed, around ten years ago, the need was felt to furnish a simplified tool for learners – Sketch Engine for Language Learning (SkELL) – which offers a reduced set of corpora and a more user-friendly interface. Kilgarriff et al. (2015, 66) describe it as "a stripped-down, non-scary version of Sketch Engine for use by learners." Moreover, many other simplified corpus resources exist, e.g., BNClab (Brezina et al. 2018), Collocaid (Frankenberg-Garcia et al. 2019) and the very recent CorpusMate (Crosthwaite and Baisa 2024). See Gerigk (2023) for a comparison of BNClab,

CorpusMate and CQPweb. In short, it could be contested that the 'scariness' of Sketch Engine makes it unsuitable for the average language learner, and that therefore the difficulties discussed above are misplaced.

This apparent paradox dissipates if we recall (i) that "teachers and students" are amongst the target users of Sketch Engine,[6] and (ii) the great lengths to which the Sketch Engine team has gone in order to make the resource and interface comprehensible and manageable for users who have not had any classroom training in the use of Sketch Engine, providing a user guide, a quick-start guide, video tutorials, face-to-face training, glossaries, interaction with the team for questions, workshops etc. Further, the user interface was recently overhauled in an attempt to render it less complex. Considering the strenuous efforts made to enhance the clarity and usability of the website, it would seem far-fetched to infer that Sketch Engine targets only those users who have had specific training.

Yet, personal experience informs me that even with training, it is very easy for learners to flounder when adopting Sketch Engine. Despite instruction, they struggle to absorb or at least recall the distinguishing features of the different query types, as a result constructing poor searches and being confounded or at least misled by the outcomes. Kennedy and Miceli (2001, 86-87) report students' reactions, for example, when no occurrences are retrieved but many are expected. Feedback included: "the phenomenon does not exist" and "the search didn't work," but in fact, as reported in the journey above, there may be thousands of occurrences. The learner, or perhaps even the teacher, may simply (i) have made misguided searches owing to the complexity of the functions available, or (ii) have been led astray by arcane messages or questionable/erroneous tagging.

7. Conclusions

In this paper it has been suggested that, for a variety of reasons, Sketch Engine queries are more complex than they seem. If relatively inexperienced users undertake the journey described above, the repeated failure to capture intelligible solutions even to fairly basic queries is likely to transform that journey into an obstacle course, and therefore to trigger such disheartenment that those users end up seeking alternative resources elsewhere. Despite the multifarious options available to Sketch Engine users to assist their searches, there is a considerable amount of information to assimilate, with the attendant risk of losing one's way in the forest of instructions and data. In short, the high degree of sophistication of this software, together with some mysterious tagging and misleading flags following unsuccessful queries, risks derailing even more experienced users. In a digital world where the increasing expectation is to be able

[6] https://www.sketchengine.eu/#blue. Last visited 06/06/2025.

to start conducting searches immediately and intuitively, the modern competition from GenAI applications is likely to be fierce.

Bionote

Dominic Stewart teaches linguistics and Italian-English translation at the Department of Humanities, University of Trento. He previously taught at the universities of Bologna, Macerata and Verona. He publishes mainly in the research areas of corpus linguistics and translation.

Works cited

Bernardini, Silvia. "Corpora in the Classroom: An Overview and some Reflections on Future Developments." *How to Use Corpora in Language Teaching*. Edited by John Sinclair. Amsterdam: John Benjamins, 2004. 15-36.

Biber, Douglas, Randi Reppen and Susan Conrad. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

Boulton, Alex. "Data-Driven Learning: In Conversation with Alex Boulton." *Corpora for Language Learning: Bridging the Research-Practice Divide*. Edited by Peter Crosthwaite. London: Routledge, 2024. 43-53.

Boulton, Alex. "Applying Data-Driven Learning to the Web." *Multiple Affordances of Language Corpora for Data-Driven Learning*. Edited by Agnieszka Leńko-Szymańska and Alex Boulton. Amsterdam: John Benjamins, 2015. 267-295.

Boulton, Alex and Nina Vyatkina. "Thirty Years of Data-Driven Learning: Taking Stock and Charting New Directions over Time." *Language Learning and Technology* 25.3 (2021): 66-89.

Boulton, Alex and Tom Cobb. "Corpus Use in Language Learning: A Meta-Analysis." *Language Learning* 67.2 (2017): 348-393.

Brezina, Vaclav, Dana Gablasova and Susan Reichelt. *BNClab*. http://corpora.lancs.ac.uk/bnclab. Last visited 1/12/2024.

Charles, Maggie. "Corpora and Autonomous Language Learning." *The Routledge Handbook of Corpora and English Language Teaching and Learning*. Edited by Reka Jablonkai and Eniko Csomay. Abingdon: Routledge, 2023. 406-419.

Crosthwaite, Peter. *Corpora for Language Learning: Bridging the Research-Practice Divide*. London: Routledge, 2024.

Crosthwaite, Peter and Lisa Cheung. *Learning the Language of Dentistry: Disciplinary Corpora in the Teaching of English for Specific Academic Purposes*. Amsterdam: John Benjamins, 2019.

Crosthwaite, Peter and Vit Baisa. “Generative AI and the End of Corpus-Assisted Data-Driven Learning? Not so Fast!” *Applied Corpus Linguistics* 3.3 (2023): 100066.

Crosthwaite, Peter and Vit Baisa. “A User-Friendly Corpus Tool for Disciplinary Data-Driven Learning: Introducing CorpusMate.” *International Journal of Corpus Linguistics* 29.4 (2024): 595-610.

Crystal, David. *A Dictionary of Linguistics and Phonetics*. Blackwell: Oxford, 2008.

Flowerdew, John. “Data-driven Learning: From *Collins Cobuild Dictionary* to ChatGPT.” *Language Teaching* (2024): 1-18.

Flowerdew, Lynne. “Data-Driven Learning and Language Learning Theories: Whither the Twain shall Meet.” *Multiple Affordances of Language Corpora for Data-Driven Learning*. Edited by Agnieszka Leńko-Szymańska and Alex Boulton. Amsterdam: John Benjamins, 2015. 15-36.

Frankenberg-Garcia, Ana, et al. “ColloCaid: A Tool to Help Academic English Writers Find the Words they Need.” *CALL and Complexity – Short Papers from EUROCALL 2019*. Edited by Fanny Meunier, et al., 2019. 144-150.

Gavioli, Laura. “Corpus Analysis and The Achievement of Learner Autonomy in Interaction.” Edited by Linda Lombardo. *Using Corpora to Learn About Language and Discourse*. Bern: Peter Lang, 2009. 39-71.

Gerigk, Kevin Frank. “Review. CQPweb, BNClab, and CorpusMate and their Applicability to the DDL Classroom.” *Árboles y Rizomas* 5.2 (2023): 144-150.

Gilquin, Gaëtanelle and Sylviane Granger. “Using Data-driven Learning in Language Teaching.” *The Routledge Handbook of Corpus Linguistics*. Edited by Anne O’Keeffe and Michael McCarthy. London: Routledge, 2022. 430-442.

Johns, Tim. “Data-Driven Learning: An Update.” *TELL&CALL* 2 (1993): 4-10.

---. “Should you be Persuaded: Two Samples of Data-Driven Learning Materials.” *English Language Research Journal* 4 (1991): 1-16.

Kennedy, Claire and Tiziana Miceli. “An Evaluation of Intermediate Students’ Approaches to Corpus Investigation.” *Language Learning and Technology* 5.3 (2001): 77-90.

Kilgarriff, Adam, et al. “Corpora and Language Learning with the Sketch Engine and SKELL.” *Revue Française de Linguistique Appliquée* 1.10 (2015): 61-80.

Knowles, Gerry and Zuraidah Mohd Don. “The Notion of a ‘Lemma’: Headwords, Roots and Lexical Sets.” *International Journal of Corpus Linguistics* 9.1 (2004): 69-81.

Leech, Geoffrey. “Teaching and Language Corpora: A Convergence.” *Teaching and Language Corpora*. Edited by Anne Wichmann, et al. Harlow: Addison Wesley Longman, 1997. 11-23.

Leńko-Szymańska, Agnieszka and Alex Boulton. “Introduction. Data-Driven Learning in Language Pedagogy.” *Multiple Affordances of Language Corpora for Data-Driven Learning*. Edited by Agnieszka Leńko-Szymańska and Alex Boulton. Amsterdam: John Benjamins, 2015. 1-14.

Moon, Rosamund. “What Can a Corpus Tell Us about Lexis?”. *The Routledge Handbook of Corpus Linguistics*. Edited by Anne O’Keefe and Michael McCarthy. London: Routledge, 2010. 197-211.

O’Keeffe, Anne. “Data-Driven Learning: A Call for a Broader Research Gaze.” *Language Teaching* 54.2 (2021): 259-272.

Zadorozhnyy, Artem and WanYeeWinsy Lai. “ChatGPT and L2 Written Communication: A Game-Changer or Just Another Tool?” *Languages* 9 (2024): 5.