Renaissance and Reformation
Renaissance et Réforme

Kilgariff, Adam, founder; Pavel Rychlý, co-founder; and Miloš Jakubíček, CEO. Sketch Engine. Other

Mel Evans

Volume 44, numéro 4, automne 2021

URI : https://id.erudit.org/iderudit/1089357ar
DOI : https://doi.org/10.33137/rr.v44i4.38650

Aller au sommaire du numéro

Éditeur(s)

Iter Press

ISSN

0034-429X (imprimé)
2293-7374 (numérique)

Découvrir la revue

Citer ce compte rendu

Evans, M. (2021). Compte rendu de [Kilgariff, Adam, founder; Pavel Rychlý, co-founder; and Miloš Jakubíček, CEO. Sketch Engine. Other]. *Renaissance and Reformation / Renaissance et Réforme*, *44*(4), 217–223. https://doi.org/10.33137/rr.v44i4.38650

Kilgariff, Adam, founder; Pavel Rychlý, co-founder; and Miloš Jakubíček, CEO.
Sketch Engine. Other.
Czech Republic: Lexical Computing. Accessed 22 July 2021.
sketchengine.eu

Sketch Engine is an online corpus query system that allows users to analyze the linguistic properties of a range of pre-loaded corpora or to explore their own corpora using a set of in-built tools. Something of a veteran in the domain of corpus linguistic tools, Sketch Engine was launched in 2004 for lexicographic applications, as part of the corpus revolution in dictionary construction. Since then, Sketch Engine has expanded in scope, accessibility, and sophistication.

Its users include academic researchers (spanning disciplines in the social sciences and humanities), professional lexicographers (including dictionary publishers such as Macmillan and Oxford University Press), and English Language Teaching (ELT) practitioners and materials designers. The system uses a commercial model, with monthly and annual subscription tariffs catering to individual academics (faculty and students), professionals, and institutions. This model has arguably ensured Sketch Engine's longevity, and likely underpins its comprehensive and professional documentation and guidance resources for users (including videos, fora, and one-to-one training options). Presently, access to Sketch Engine is free for users affiliated with academic institutions in the EU and the UK until March 2022, a scheme generously funded by ELEXIS (elex.is).

The relevance of Sketch Engine for early modernists interested in digital approaches is not immediately obvious. Of the five hundred plus corpora preloaded on the system, only a few currently relate to the early modern period (see Table 1). The English Historical Book Collection, collecting EEBO, EECO, and Evans materials in a corpus of over eight hundred million words, is the most substantial resource, and its constituent parts will be familiar to many readers who may have used the TCP files via other platforms. The Sketch Engine interface provides a different perspective to these other points of access (i.e., the Gale Cengage search pages). Sketch Engine provides users with a set of tools that can analyze early modern texts and datasets without requiring detailed knowledge of code or database systems. Moreover, users are able to create their own corpora to use in the system alongside the pre-loaded corpora.

Table 1: Pre-loaded historical corpora on Sketch Engine

Corpus of English Dialogues	Speech-related texts created between 1560 and 1760, comprising trial proceedings, dramatic comedy, didactic works, witness depositions, and prose fiction. 1.2 million words. Access provided upon special request.
Penn Corpora of Historical English	Texts created between 1150 and 1950. Includes the Penn-Helsinki Parsed Corpus of Middle English, Early Modern English, and Modern British English, which collect samples of English texts across multiple genres over time. Plain text, part-of-speech, and syntactically parsed versions available. 3.8 million words. Access provided upon special request.
GerManC	Text samples from German newspapers ca. 1650–1800. Samples are around two hundred words, with detailed metadata and division into fifty-year sub-periods. 667,000 words. Annotated using TreeTagger, with part-of-speech tagset adapted for early modern German.
English Historical Book Collection	Texts published between 1473 and 1820: includes EEBO-TCP phase I (25,364 texts, 1473–1700), ECCO-TCP (2473 texts, 1701–1800), and Evans (5007 texts, 1639–1800). 826 million words. Annotated using Penn TreeBank tagset. Not accessible to free trial users.
Transhistorical Corpus of Written English	Texts created between 1500 and 2020. Five genres: sermons, statutes, letters, emails and instant messages (the last two are restricted to late twentieth century onwards). The corpus draws on existing collections for its historical material, such as Corpus of Early English Correspondence. 501,000 words. Modern texts annotated using TreeTagger.
Latin Corpus	Collection of Latin texts from classical to medieval periods, spanning literary, historical, philosophical, and poetic text-types. 11.2 million words. Annotated using TreeTagger trained for historical Latin.

Sketch Engine offers nine tools that provide rich linguistic information about the language in a given corpus. Word Sketch, after which the system is named, provides an overview of the grammatical and lexical properties of a given search term, showing the top co-occurring words (collocates) in different syntactic positions. Figure 1 shows part of the word sketch for the address term "sir" in the English Historical Book Collection. The word sketch is interactive, and users can explore specific collocates, e.g., "sir pray you," in more detail in the corpus. There is also a visualization tool to summarize these results. A related tool, Word Sketch Difference, allows users to compare word sketches across different corpora.

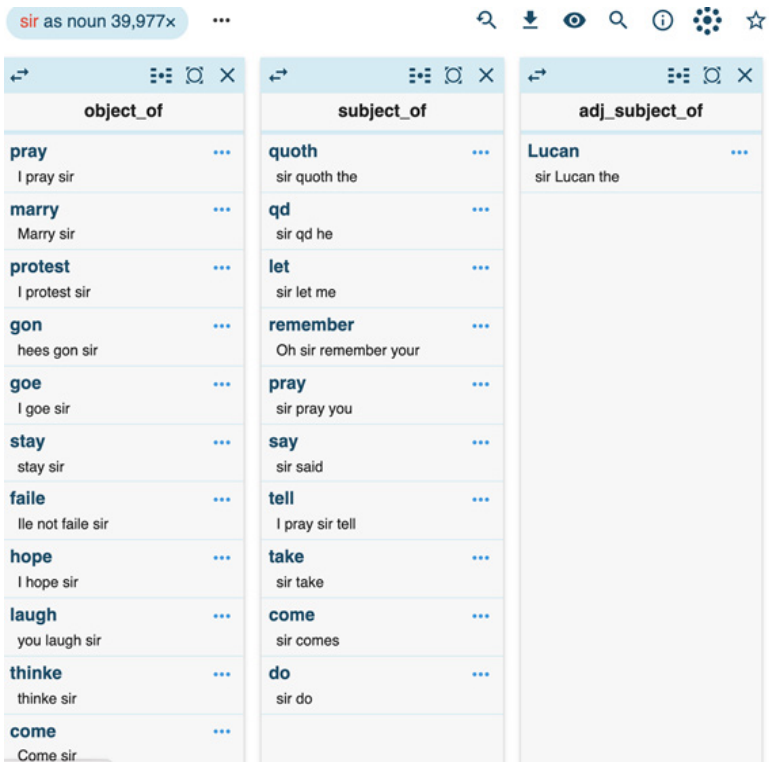

Figure 1. Screenshot of Word Sketch for "sir" in the Historical English Books Collection.

Other tools include the Thesaurus tool, which shows the frequency of synonyms or items in a semantic field of a given search term. This relies on the corpus using a standardized tagset, and best results are dependent upon high-frequency items. The Concordance tool identifies the top collocates (co-occurring words) of a search term and can be filtered by criteria such as text-type, date, and sub-section of document, as well as displaying the distribution across a corpus and creating random samples of concordance lines for closer analysis. The parallel concordance tool has similar functionality but works across two corpora in different languages. Word List produces a frequency-ranked list of all lexical items in a corpus, and N-Grams identifies sequences of tokens (e.g., words, lemmas) that can be assessed based on frequency. The Keywords tool ranks lexis that is statistically more frequent in the target corpus compared to a reference corpus, and typically offers insights into the topicality and themes of a corpus. The final tool, Trends, relies on a corpus that has temporally tagged information; for the Historical Book Collection, for example, users can analyze the frequencies of words in the corpus by decade, revealing their main trajectories (increase or decrease) over time.

A distinctive feature of the Sketch Engine system is the ability for users to upload their own corpora, and to apply the sophisticated tools through the user interface to their data. Corpora can be compiled from files uploaded directly to Sketch Engine, or created by scraping the web, using targeted search terms, URLs, or whole websites. This option has clear opportunities for early modernists working with newly digitized materials, for example, or manuscript texts such as correspondence. With appropriate preparation (discussed below), user corpora can be contrasted with other corpora in Sketch Engine.

The Sketch Engine tools therefore provide a valuable set of interlinked perspectives on the language of a corpus. Rather than surmise facts about usage based on erratic search results in EEBO or Literature Online, for example, the Sketch Engine interface produces robust, statistically informed, and replicable information about the language of early modern texts. Users interested in collocational behaviour of topic-related lexis, for instance, can explore keywords in context using the keywords and concordance tools. The distribution of words can be tracked over time and across genres using the trends tool. And the Word Sketch Difference tool can reveal important distinctions surrounding lexical meanings (and intrinsic ideologies) by contrasting genre- or text-types. These tools can be profitable for scholars working across literary and historical

disciplines, whether investigating the history of ideas or testing hypotheses of authorial style.

One of Sketch Engine's strengths is its intuitive and slick user interface. The main access point is the Dashboard, which displays the main tools and the corpus currently in use, as well as a list of recently used corpora, recent searches, and any annotations made to user corpora. To get started, users must first select their corpus (whether preloaded or the user's own). Once selected, this corpus will be loaded for use across all analytic tools, with the name displayed, for reference, in the central top header. This useful feature, like many on the platform, reflects the years of refinement to the interface design in identifying what is most practical and helpful. All tools can be accessed at any time via an overlaid left-hand sidebar, with the selected tool then appearing in the main window. The outputs from the corpus analysis are shown on screen, with different kinds of interactivity possible (e.g., clicking on a search term to view it in context). Some tools, such as Word Sketch Difference, have in-built visualizations that can be saved by the user.

Outputs from corpus analyses can be downloaded in multiple file formats (e.g., xml, csv, PDF) for use in other applications. While downloads of search results and analysis of the users' own corpora are unrestricted, the system caps download capacity for the pre-loaded corpora: e.g., for trial and ELEXIS accounts, this is one thousand collocations and/or ten thousand concordance lines from one concordance. This reflects the licensing requirements of the system, which effectively hosts corpora each with their own restrictions of use on copyrighted material. There is also a standard cap on the storage space granted to each user for their own corpora (one million words). For academic researchers, these restrictions can be removed upon request, subject to the approval of the Sketch Engine operators. That said, the quantities given are generous in the context of early modern materials and would likely be sufficient for many users.

The platform also accommodates different skill levels. While the interface by default allows all essential tools to be automated by the system, the user can also access advanced options for searches, corpus annotation, and analytic tools to produce a more curated and specialized set of results. This includes a JSON format API, governed by a Fair Use Policy. As commercial software, the Sketch Engine documentation is comprehensive. This includes advanced documentation for expert users, documentation for the API and Fair Use Policy, and a general user's guide targeted at non-specialist users.

With its effective interface and capacity to accommodate different user needs and skill levels, Sketch Engine is a well-developed piece of software. Underlying the system is Manatee, a corpus linguistic specific database management system, where sequence, not relations, is the connecting principle. Corpus Query Language (CQL) is used to run queries of the relevant corpus, producing word sketches and related outputs. The interface uses Bonito at the front end, and the Corpus Architect module builds and manages user corpora. Within the corpus-side of the system, users can annotate their own corpora to mark structure and provide content metadata. For the former, users are encouraged to apply only simple XML mark-up relating to core structural facets, i.e., documents, paragraph, sentence, and more complex tagging may not be processed accurately by the module. Corpora can also be annotated using standardized tagsets, such as Penn TreeBank, to demarcate parts of speech. Tagsets are available for over thirty languages. For user corpora, metadata can be added using an online form which automatically inserts the information into the header. Users are not, for obvious reasons, able to modify the metadata or annotations of the pre-loaded corpora.

Despite being developed for present-day lexicography, Sketch Engine has much potential for scholars interested in digital approaches to early modern studies. Corpus linguistic techniques are expanding beyond their original language remit, and the potential of what elsewhere might be called distant reading is being realized across the humanities, arts, and social sciences. Past work using Sketch Engine with early modern texts includes the analysis of colonial ideologies in British pamphlets on Virginia[1] and specialized vocabulary in early modern translations,[2] as well as language-focused studies of pragmatic, lexical, and syntactic features across time and genres. As interdisciplinary ways of working continue to inform early modern studies, tools like Sketch Engine will become a valuable resource for researchers' investigation of complex questions about languages and texts and the individuals and societies that produced them.

MEL EVANS
University of Leeds
https://doi.org/10.33137/rr.v44i4.38650

1. Cecconi.

2. Ciambella.

Works Cited

Cecconi, Elizabeth. 2020. "British Colonial Ideology in the Language of Pamphlets on Virginia (1584–1624)." In *The Language of Discovery, Exploration and Settlement*, edited by Nicholas Brownlees, 8–24. Newcastle Upon Tyne: Cambridge Scholars Publishing.

Ciambella, Fabio. 2019. "Specialised Tourist Discourse: Translating Schott's *Itinerarii* in Early Modern England." In "Translation and the Non-literary Text: From Early to Late Modern English." Special issue, *Status Quaestionis* 17.

Eder, Marciej, Jan Rybicki, and Mike Kestemont, project creators. Stylo(). Other.
Krakow: University of Kracow, Poland. Accessed 17 February 2022. github.com/computationalstylistics/stylo.

R Stylo is a suite of programs, the most widely used of which is stylo(). Stylo() enables stylistic exploration of lexical corpora. It serves as a testbed for wide-ranging experimentation with texts of varying size. It enables the use of a multiplicity of parameters that scrutinize the authorship of early modern texts. As evidence of its flexibility, five settings of three parameters allows for 3^5, or 243, possibilities, more than is practical with manual operation. And there are far more than five choices of three parameters in stylo().

One appreciates the advantages of stylo() as a tool, by comparing its simplicity with the multiple stages ordinarily required for statistical stylistic analysis. For an early modern play, one selects reliable electronic texts in original or standardized modern spelling, with their speech headings and stage directions commonly removed. This preliminary stage is the same with or without stylo().

The ensuing stages, however, are the ones in which R Stylo excels. Stylo() comes with a graphical user interface (GUI) which enables a researcher without programming skills to regulate its parameters. To interrogate, for instance, the authorship of a disputed early modern play, one selects a set of linguistic parameters. Stylo() automatically selects the most frequent variables determined by those pre-chosen parameters. These include the number of most