

Ambientes tecnológicos para o estudo da linguagem – destaque para a plataforma Sketch Engine

Technological spaces for language studies – focus on Sketch Engine

Eloá Bera Santana¹

Regiani Aparecida Santos Zacarias²

Resumo: O presente artigo resulta de pesquisa que avaliou a plataforma digital Sketch Engine (SKE) quanto ao seu diferencial face às demais ferramentas de igual natureza. Da mesma forma, verificou as funcionalidades dessa plataforma, de suas ferramentas e extensões para aplicação aos estudos da linguagem e à elaboração de dicionários bilíngues. A metodologia teve base investigativa e descritiva, com abordagem qualitativa e combinou revisão bibliográfica, estudo e exploração prática da SKE, visando exemplificar seu uso em contextos autênticos. O objetivo deste artigo, portanto, é apresentar as principais características da SKE que a distinguem de outras plataformas da mesma natureza e descrever suas ferramentas e recursos, exemplificando, com ênfase, a funcionalidade da ferramenta *Parallel Concordance* aplicável aos estudos linguísticos bilíngues voltados à tradução e à lexicografia. A fundamentação teórica selecionada é baseada nos princípios e conceitos da Linguística de *corpus*, da Lexicografia Geral e da Lexicografia digital. Devido ao dinamismo com que os modelos computacionais têm avançado, fomentar conhecimento sobre a SKE em língua portuguesa brasileira para ampliar as suas possibilidades no âmbito dos estudos linguísticos de base em *corpora* é imprescindível.

Palavras-chave: E-dicionário. Lexicografia. *Corpus* linguístico.

Abstract: This article reports on research that assessed the Sketch Engine digital platform (SKE) with regard to its distinctive features in comparison with similar tools. It also examines the platform's functionalities and extensions for application in language studies and the development of bilingual dictionaries. The methodology was investigative and descriptive, adopting a qualitative approach that combined a literature review with practical exploration of SKE to demonstrate its use in authentic contexts. The aim of this paper is to present the main characteristics that distinguish SKE from other platforms and to describe its resources, with particular emphasis on the *Parallel Concordance* tool and its applicability to bilingual studies focused on translation and lexicography. The theoretical framework is grounded in the principles and concepts of Corpus Linguistics, General Lexicography, and Digital Lexicography. Given the rapid advancement of computational models, promoting knowledge of SKE in Brazilian Portuguese is essential to expand its potential in corpus-based linguistic studies.

Keywords: E-dictionary. Lexicography. Linguistic *corpus*.

¹ Mestranda na Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Exatas, Departamento de Letras Modernas, Programa de Pós-Graduação em Estudos Linguísticos, São José do Rio Preto, SP, Brasil. Endereço eletrônico: eb.santana@unesp.br.

² Professora Livre-Docente na Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências, Departamento de Educação, Programa de Pós-graduação em nível de Mestrado Profissional em Docência para Educação Básica, Bauru, SP, Brasil. Endereço eletrônico: regiani.zacarias@unesp.br.

Introdução

Este trabalho³ tem como temática a exploração e a expansão dos estudos envolvendo a Linguística de *Corpus*, a Linguística Computacional e a Lexicografia digital eletrônica (também denominada e-lexicografia). O artigo revela inovadoras perspectivas ao estudo da linguagem para fins lexicográficos, oportunizadas por novos meios de compilação e tratamento de *corpora* e uso de plataformas fundamentadas em modelos de Inteligência Artificial. Entende-se que, com a dinamicidade com que os modelos computacionais têm avançado, torna-se necessário compreender sistemas e estruturas que possam atribuir acurácia e aprofundamento aos estudos no campo da língua e da linguagem. Sendo assim, esta proposta visa a colaborar com a demanda existente por literatura teórico-prática sobre e-lexicografia no Brasil.

Para colaborar com o atendimento a essa demanda, este estudo avaliou a plataforma digital Sketch Engine (SKE) quanto ao seu diferencial face às demais plataformas de mesma natureza, às funcionalidades de suas ferramentas, às extensões para aplicação em pesquisas voltadas aos estudos da linguagem e à elaboração de dicionários bilíngues. A escolha da SKE como objeto de estudo deu-se por ser a única de sua natureza, especificamente voltada para atender à lexicografia digital/eletrônica (e-lexicografia). A Sketch Engine é um exemplo de software de análise de *corpus* e texto, composto por um website com ferramentas principais e secundárias para analisar vários *corpora* simultaneamente e desenvolver estudos linguísticos, lexicográficos e de tradução (Kilgarrieff, 2014). Fundada em 2003, por meio da parceria do linguista Adam Kilgarrieff e do cientista de computador Pavel Rychlý, a SKE faz uso de tecnologias tais quais a de processamento por algoritmo, consolidando-se um recurso computacional dotado de inúmeras possibilidades e fecundo campo de estudo, trabalho e pesquisa. É um expoente entre diversos modelos digitais elaborados por meio da Ciência da Computação, sobretudo, o de seu sub-ramo mais frutífero: a Inteligência Artificial (IA). Nesse caso, por se tratar de uma plataforma baseada em *corpus*, também usufrui da ciência de dados para construir *corpora*, dispor os dados on-line e analisá-los com recursos lexicográficos

³ Pesquisa de Iniciação Científica financiada pela FAPESP, Proc. nº. 2025/06136-9, intitulada “Lexicografia Computacional: Corpora e inteligência artificial na elaboração de dicionários digitais”.

reproduzidos por modelos de linguagem potencializados pelos estudos de *machine learning* e, mais recentemente, por Inteligência Artificial.

Nesse contexto, a proposta deste artigo é apresentar as principais características da SKE que a distinguem de outras plataformas da mesma natureza e descrever suas ferramentas e recursos, exemplificando a funcionalidade da ferramenta *Parallel Concordance*, aplicável aos estudos linguísticos bilíngues voltados à tradução e à lexicografia.

Fundamentação Teórica

A bibliografia selecionada discorre sobre os princípios teóricos da Linguística de *corpus*, da Lexicografia Geral e da Lexicografia digital, contemplando autores como Sidney I. Landau, Vander Viana, Sylviane Granger e Gilles-Maurice de Schryver, além dos registros dos idealizadores do Sketch Engine, Adam Kilgarriff e Pavel Rychlý. Através dessas leituras, procurou-se mapear e compreender como as teorias podem ser observadas no uso da SKE. Também houve contribuições de exemplos de *corpora* para ilustrar e complementar as análises bibliográficas.

Uma das mais recentes disciplinas linguísticas, a Linguística de *corpus*, tem, desde o seu surgimento, impactado a lexicografia no tocante à elaboração e ao uso de dicionários e outras obras de referência. Viana (2008, p. 31) define a Linguística de *Corpus* como uma compilação de textos que viabiliza pesquisas no campo da Linguística:

[...] uma compilação eletrônica e criteriosa de (amostras de) textos que ocorrem naturalmente com o objetivo de representar uma dada língua ou algum de seus aspectos mais pontuais de forma a possibilitar uma análise linguística previamente delineada (Viana, 2008, p. 31).

Viana (2011) esclarece ainda que *corpus* é um conjunto de textos orais ou escritos, efetivamente produzidos por falantes de determinada língua, de modo a representar empiricamente o uso dela em contextos gerais e específicos. A partir do estabelecimento de critérios bem definidos, o *corpus* constitui uma amostra representativa de uma dada população em dado contexto linguístico.

Os estudos linguísticos com base em *corpora* permitem a análise de características e padrões linguísticos. Nesse contexto, é possível a extração de lemas, fraseologias, estruturas sintáticas e sintagmáticas para gerar resultados a serem observados em suas repetições e construções de sentidos, dentre outros aspectos.

A Linguística de *Corpus* contribui, ainda, tanto para pesquisas linguísticas quanto literárias, com análises que podem colaborar para a estilística literária, os estudos de tradução, a preservação linguística, as equivalências ou diferenças entre vários idiomas e outras possibilidades infinitas a depender da visão do linguista. Tudo isso por meio da compilação, seleção e análise de *corpora* – atividades potencializadas por meio do trabalho lexicográfico, produzido por lexicógrafos.

Por ser contemporânea às evoluções tecnológicas ocorridas, sobretudo, a partir da primeira e da segunda década de 2000, a Linguística de Corpus beneficia-se dos avanços tecnológicos que permitiram, primeiramente, a migração de material linguístico impresso para o meio eletrônico e, atualmente, a criação de material linguístico em formato digital eletrônico, expandindo o armazenamento de dados linguísticos, promovendo eficácia e confiabilidade na extração de resultados.

A Linguística de Corpus também passou a ocupar-se de ferramentas para consultas aos *corpora*. Surgiram sistemas que são dotados de linguagem computacional em algoritmos e que, ao serem alimentados por *corpora*, permitem buscas de dados a partir das opções disponíveis pelo próprio sistema. A esse respeito, Höfling e Zacarias (2017) esclarecem as possibilidades de recursos (*corpus query system*) para a consulta a *corpora*, oferecidas pelo *WordSmith Tools*:

[...] o programa *WordSmith Tools* oferece as ferramentas Concordanciador (*Concord*) para a busca de concordâncias; palavras-chaves (*Keywords*) a partir de critérios estatísticos pré-determinados; e lista de palavras (*Wordlist*) em um *corpus*, como principais” (Höfling; Zacarias, 2017, p.112).

As autoras explicam que o *WordSmith Tools* é um tipo de sistema de busca de dados linguísticos que não hospeda *corpora*; por isso, para usá-lo é necessário compilar um *corpus* digitalizado, fazer o upload e aplicar as ferramentas aos textos.

Certamente, o avanço das tecnologias digitais foi um marco e um salto nos estudos de base linguística. Mais recentemente, a Linguística Computacional desenvolveu sistemas

com algoritmos criados a partir de estudos das redes neurais que possibilitaram o cruzamento de dados linguísticos, ou seja, por meio desses, os sistemas computacionais passaram a 'pensar'. Posteriormente, a chamada Inteligência Artificial (IA) foi desenvolvida e abriu caminhos para novas possibilidades de pesquisas, registros e geração de dados.

A Lexicografia, ciência que *stricto sensu* se ocupa da elaboração de dicionários, tem seguido a sua trajetória de registros linguísticos concomitantemente com os avanços tecnológicos. Ao final da década de 70, a Lexicografia passou a valer-se de *corpora* linguísticos como fonte para a criação de dicionários e, atualmente, beneficia-se da inteligência artificial para dinamizar o processo e inovar a interface que disponibiliza aos usuários.

Para entender por que a tecnologia favorece tanto o processo de criação dos dicionários, é preciso recorrer às suas origens, conforme os autores clássicos definem:

Um dicionário é um texto que registra as palavras de uma língua e explica seus significados, ou fornece palavras equivalentes em outro idioma, geralmente organizado em ordem alfabética⁴ (Landau, 2001, p. 7, tradução nossa).

Dicionários são criados para serem consultados em diversos tipos de atividades, como as de escrita, as de leitura e as complexas que exigem atenção do sujeito envolvido no processo de aprendizagem. Também podem simplesmente ser consultados para sanar dúvidas pontuais e curiosidades daqueles que se interessam em saber mais sobre as palavras das línguas. Desde o seu surgimento e até o final do século XX, as listas de palavras, que antecedem os dicionários, eram um material lexicográfico cuja compilação era realizada manualmente. Esses procedimentos estendiam-se por anos até a finalização e a revisão das obras.

Dessa forma, como afirma Robert Lew (2024), os trabalhos lexicográficos costumavam ser atividades conservadoras, firmadas em convenções e cânones, e o conteúdo de tais obras consistia em apresentar definições, sinonímia, etimologia e/ou categorias gramaticais. Sustentados, em geral, por estruturas semasiológicas, que orientavam as pesquisas da forma para o significado, os dicionários monolíngues foram

⁴ “A dictionary is a text that records the words of a language and explains their meanings, or gives equivalent words in another language, usually arranged in alphabetical order.”

costumeiramente considerados mantenedores da língua que registravam, enquanto os bilíngues serviam de apoio em pesquisas, práticas de tradução e aprendizagem de idiomas, por promoverem suporte lexical para compreensão e produção de textos.

Alguns aspectos lexicológicos são complexos quando estudados e abordados para fins lexicográficos, como a identificação e a catalogação de fraseologias, linguagem conotativa e expressões idiomáticas, equivalências de palavras e de fraseologias; se realizados manualmente, podem levar muito tempo e ainda, em geral, carecem de precisão. Nesse sentido, o uso das tecnologias digitais foi um marco e um salto nos estudos lexicográficos, o registro de *corpora* em redes neurais, os estudos de usuário e as referências cruzadas proporcionaram um novo tipo de pesquisa e registro de dados.

A velocidade das ferramentas digitais e, posteriormente, daquelas com inteligência artificial, tornou o trabalho lexicográfico mais ágil e produtivo. Esse novo modo de trabalhar a lexicografia e de disponibilizar as obras em meio eletrônico tem sido conhecido como *e-lexicografia*⁵. A linguista Sylviane Granger, em “Lexicografia eletrônica: de desafio à oportunidade” (2012), introduz e explica como os dicionários migraram para o meio digital e tornaram-se instrumentos do que hoje é conhecido como *e-lexicografia*:

Hoje lexicografia é amplamente sinônima a lexicografia eletrônica e muitos especialistas predizem o desaparecimento dos dicionários de papel em um futuro próximo [...] lexicografia eletrônica é um termo guarda-chuva para referir-se ao design, uso e aplicação de dicionários eletrônicos (DEs), que definem-se como coleções primordialmente humanizadas de dados eletrônicos estruturais que fornecem informações sobre forma, significado e uso de palavras em um ou mais idiomas e são armazenadas em uma gama de dispositivos (PC, internet, dispositivos móveis).⁶ (Granger, 2012, p. 2-3, tradução nossa).

A capacidade de retenção de informação digital também é expandida de forma significativa, com a oportunidade de extrair e armazenar dados linguísticos em maior escala, com muita rapidez, e estruturá-los conforme sua finalidade, seja científica, seja

⁵ Termo utilizado por renomados autores como Fuertes-Oliveira, P.A. e Bergenholtz H. (eds.). *e-Lexicography*. London, New York: Continuum, 2010. Também presente no website *European Network of e-Lexicography* - <http://www.elexicography.eu/>.

⁶ “Today lexicography is largely synonymous with electronic lexicography and many specialists predict the disappearance of paper dictionaries in the near future. [...] electronic lexicography is used as an umbrella term to refer to the design, use and application of electronic dictionaries (EDs), which are in turn defined as primarily human-oriented collections of structured electronic data that give information about the form, meaning and use of words in one or more languages and are stored in a range of devices (PC, internet, mobile devices)”

pedagógica, ou recurso para tradução. Por isso, muitos cientistas como Sylviane Granger (2012) apontam para a possível extinção de dicionários físicos e especula-se o fim da profissão dos lexicógrafos em um futuro próximo, devido à nova potência tecnológica dos anos 20: a inteligência artificial. Contudo, tendo em vista a análise dos recursos existentes e suas possibilidades, compreende-se que a automatização do trabalho de compilação e corpora não exclui o fator humano das pesquisas lexicográficas, mas passa a exigir profissionais multifacetados, com habilidades tanto para a ciência da computação, quanto para as ciências linguísticas.

Metodologia

A pesquisa da qual deriva este artigo é de natureza investigativa e descritiva, examinando e descrevendo a SKE como recurso para a geração de dados linguísticos para fins lexicográficos e para a criação de dicionários digitais. Quanto à abordagem do problema, classifica-se como qualitativa porque procede à análise do objeto de estudo, no caso, a SKE.

Esta metodologia foi selecionada pois por meio de seus processos atende às necessidades da pesquisa de forma adequada. Para a análise da SKE, utilizou-se a coleta de dados transversal, através do acesso pago e gratuito dos *corpora* disponíveis na plataforma, explorando também suas ferramentas, tais quais *Word Sketch*, *Collocations* e *Concordance*, com o intuito de identificar e descrever suas funcionalidades para estudos lexicográficos, com para falantes de Língua Portuguesa.

Resultados e discussão

1.1. Recursos computacionais para estudos linguísticos

Para entender a qualidade da Plataforma Sketch Engine, uma pesquisa preliminar avaliou, de forma descritiva-comparativa, a Plataforma Sketch Engine (SKE) e a comparou com plataformas de funcionalidades semelhantes: *Wordsmith Tools*, *AntConc*, *CLARIN ERIC* e *DeepL Pro*. Buscou-se comparar os limites linguísticos e tecnológicos possibilitados por cada plataforma, considerando a tecnologia disponível, a interface e a acessibilidade em relação ao público em questão, a saber, os lexicógrafos. Para tal, descreveu-se cada

uma das plataformas mencionadas, e teceram-se, a modo de discussão, comentários comparativos em relação a essas e à SKE.

1.1.2 Sketch Engine

A SKE destaca-se por ser a única plataforma disponível na web, para fins lexicográficos, que disponibiliza mais de 800 *corpora* pré-carregados e etiquetados, prontos para pesquisas, abrangendo mais de 100 idiomas. Ainda, vale-se da linguística de *corpus* e de tecnologias como o processamento de linguagem natural e de Inteligência Artificial (IA) – modelos de linguagem e *word embeddings*, especificamente – aliados a algoritmos e a estatísticas linguísticas no processamento de seus dados. Trata-se de um software capaz de fornecer um conjunto de recursos para estudos do léxico concentrados num mesmo ambiente digital, o que proporciona agilidade, organização e versatilidade para seus usuários. A diferença entre a SKE e as demais ferramentas que permitem a extração de dados em *corpora* é o fato de a SKE especializar-se no atendimento ao fazer lexicográfico; e, por essa razão, criar funcionalidades para esse fim.

Quanto aos *corpora* da SKE, são robustos e alguns possuem até cerca de 80 bilhões de palavras (Kilgariff, 2014). Além disso, os *corpora* são compilados com conteúdos variados, como legendas, artigos, blogs, documentos e outros recursos disponíveis publicamente na Internet. No âmbito bilíngue, a principal inovação é a presença dos *corpora OpenSubtitles*, compilados a partir de legendas de filmes geradas por IA e, por essa razão, de domínio público, que permitem pesquisas em *corpora* paralelos e identificação de padrões tradutórios de linguagem contemporânea das línguas modernas. Antes do surgimento dos *corpora OpenSubtitle*, em 2023, as pesquisas de padrões bilíngues ou multilíngues restringiam-se a *corpora* compilados com as obras antigas e clássicas e a documentos de domínio público, sendo também possível a compilação de *corpora* específicos e de outras fontes, por meio da solicitação da concessão de direitos autorais. Os *corpora OpenSubtitles* representaram uma revolução para a linguística contemporânea, exploráveis por meio das ferramentas “*Parallel Concordance*” e “*Bilingual Terms*”. É viável trabalhar com dois ou mais *corpora* simultaneamente em ferramentas de comparação, como o *Parallel Concordance*, por exemplo. Além dos *corpora* disponíveis, existe a opção para inserir, criar e editar *corpus* próprio.

Na SKE existem ainda *corpora* extraídos de dicionários digitais tradicionais que estão sempre atualizados e que, por serem diacrônicos, podem ser consultados pela ferramenta “*Trends*” (tendências) para gerar dados que permitam verificar o uso corrente das línguas.

Quanto aos recursos, na página inicial e principal, localiza-se uma interface (*Dashboard*) com as principais ferramentas: *Word Sketch*, *Word Sketch Difference*, *Thesaurus*, *Concordance*, *Parallel Concordance*, *Wordlist*, *N-grams*, *Keywords*, *Trends*, *Text type analysis*, *Bilingual Terms* e *One Click Dictionary* (descritas em 2.2.1). No interior das buscas dessas ferramentas principais, outras ferramentas, às quais chama-se aqui de secundárias, são disponibilizadas para aprimorar ou facilitar a pesquisa no canto superior direito. Ao gerar dados pela *Word Sketch*, por exemplo, surgem as seguintes possibilidades: mudar o critério (*change criteria*), download, mudança de layout (*change view options*), filtragem de colocados (*filter results*), informações sobre a informação apresentada (*screen details*), visualização gráfica (*visualization*), salvar como favoritos (*add to favorites*). As ferramentas principais e secundárias se completam, e cada dado gerado vem acompanhado de três pontos (...), à direita, com mais opções de pesquisas.

Uma ferramenta secundária que aprimora a busca de forma significativa é o *Corpus Query Language* (CQL – lupa superior que permite selecionar o *corpus*), serviço que possibilita pesquisar, extrair e analisar padrões de grandes quantidades de *corpora*, buscando tanto informações simples quanto estruturas complexas, pesquisando por meio de uma triagem baseada em anotações multinível (lemas, categorias gramaticais e traços morfossintáticos), encontrando padrões linguísticos com velocidade. Atua em *corpora* previamente anotados e consegue detectar recorrências lexicais, isolando estruturas para futuras análises de concordância e de frequência.

Dentre as ferramentas secundárias, também se destaca o GDEX (*Good Dictionary Examples*, ou “bons exemplos para dicionário”), uma ferramenta atrelada às funções de concordância (*Concordance* e *Parallel Concordance*), que atua como um filtro de dados, fornecendo uma seleção dos exemplos de colocações automaticamente avaliadas como sendo as mais adequadas para exemplos em contextos de aprendizagem. O usuário indica o número de sentenças que deseja, sendo o máximo de trezentas; e, com base no tamanho, na quantidade de palavras ditas ‘complexas’ utilizadas no vocabulário, no tema

(se é controverso, como, por exemplo, religião ou política), se há contexto lexical e linguístico suficiente ou há referências a algum tópico ou sentença (como pronomes com referência externa) ou detalhes como nomes de marca e critérios similares. O GDEX filtra as sentenças e indica os melhores exemplos.

A plataforma mantém o histórico dos *corpora* utilizados anteriormente e das pesquisas realizadas na conta do usuário, permitindo efetuar anotações e salvar buscas em favoritos para exibição rápida. A tecnologia de ponta que fundamenta a plataforma, unida à competência e dedicação profissional de seus criadores e mantenedores, fez com que a SKE fosse superior e singular dentre as plataformas com funções similares, como as que são descritas a seguir.

Uma característica marcante da SKE é a relativa aos tutoriais detalhados e em vídeo a respeito de toda a sua estrutura e funcionalidade, permitindo o seu uso com qualidade, mesmo por usuários iniciantes ou com pouca familiaridade no universo lexicográfico computacional. O empenho da equipe oportuniza, além de dados de pesquisa, letramento lexicográfico computacional e colabora para a democratização dessa área do saber. Nesse contexto, a SKE tem contribuído positivamente para a ponte entre a e-Lexicografia e a Linguística de *Corpus* (Kilgariff & Kosem, 2012).

A SKE permite o uso gratuito por 1 mês (*free trial*) e depois requer o pagamento, estipulado de acordo com o tipo de uso (acadêmico, profissional ou empresarial). Para instituições de ensino, há a possibilidade de convênios, e, para grupos de pesquisa, há descontos para estudantes mediante a assinatura do pesquisador principal.

Embora não haja dúvida acerca da contribuição que todas as plataformas aqui avaliadas prestem aos estudos linguísticos, a SKE destacou-se por sua estrutura abrangente e especificamente lexicográfica e por fornecer *corpora* prontos para pesquisas em mais de 100 idiomas. A análise confirmou, portanto, a relevância da SKE para os estudos da linguagem e a demanda por pesquisas que possam incluir a SKE na literatura lexicográfica computacional em língua portuguesa.

A partir dessas informações, observam-se as demais plataformas selecionadas para fim descritivo e comparativo.

1.1.3 Wordsmith Tools

A plataforma *Wordsmith Tools* (Scott, 2024) é amplamente utilizada em pesquisas linguísticas e lexicográficas desde a década de 90, ano de sua criação. É um software clássico de linguística de *corpus* que disponibiliza as ferramentas como concordanciador, listas de frequência e *keywords* (palavras-chave). Os principais usuários dessa plataforma são profissionais envolvidos em análise textual – professores, pesquisadores e estudantes de linguística, o que tornou a *Wordsmith Tools* uma referência no meio acadêmico-científico, visto sua robustez metodológica e seus métodos tradicionais e replicáveis. Apesar disso, esse software possui menos recursos integrados em comparação com o *Sketch Engine*, exigindo maior intervenção manual por parte dos usuários e preparação prévia dos corpora, o que torna o processo de análise e edição de *corpus* moroso e disperso.

1.1.4 AntiConc

O software de uso gratuito *AntConc* (Anthony, 2019) é utilizado no contexto de ensino e pesquisa, contando com os mesmos tipos de recurso do *Wordsmith Tools* para análise qualitativa dos dados. Por não ter custo, torna-se uma opção acessível para estudantes e pesquisadores em formação. Sua fonte de dados é alimentada por meio de busca textual e cálculos estatísticos, fomentando pesquisas de linguística com análises textuais. Contudo, todo o processo necessita de alimentação e gerenciamento manual de *corpora*, sendo, portanto, lento e limitado, se comparado à SKE.

1.1.5 CLARIN ERIC

O *CLARIN ERIC – Common Language Resources and Technology Infrastructure* (CLARIN ERIC, [s.d.]) – é uma infraestrutura de pesquisa criada na Europa e funciona como um ecossistema integrado de recursos linguísticos, *corpora* e ferramentas computacionais. Está voltado para fins acadêmicos e também profissionais. Conta com repositórios de *corpora* anotados, ferramentas de busca e anotação automática em diversos idiomas. Contudo, o *CLARIN ERIC* é indicado para pesquisas de base metacomputacional em humanidades digitais e linguística computacional, exigindo conhecimento técnico avançado, diferentemente do software SKE, cuja ênfase está na atividade lexicográfica. Os

dois recursos servem a públicos diferentes; o primeiro, ao interessado em pesquisas computacionais e o segundo, ao pesquisador linguista, muitas vezes, com pouco conhecimento de linguagem e conhecimento técnico computacional.

1.1.6 DeepL Pro

A *DeepL Pro* (DeepL, 2024) é uma plataforma de tradução automática, diferente das anteriores, que são ferramentas de exploração de *corpora*. Utiliza modelos de linguagem baseados em inteligência artificial e redes neurais para realizar traduções até mesmo sensíveis ao contexto. A plataforma é direcionada ao uso profissional por tradutores, pesquisadores, docentes e estudantes, além de empresas e interessados. Em sentido estrito, não permite a exploração de *corpora*, restringindo-se a dados bilíngues, permitindo encontrar equivalentes lexicais e frasais, fornecendo quantidade satisfatória que favorece o reconhecimento de padrões tradutórios. Quando observada em comparação com a SKE, a diferença entre ambas fica evidente, pois a *DeepL Pro* não disponibiliza *corpora* para pesquisas. Conclui-se, assim, que são plataformas que atendem a funções e propósitos diferentes.

Com base neste estudo preliminar, é possível inferir que a SKE se destaca por ser a única plataforma destinada às pesquisas e às produções lexicográficas. Isso significa que não existe outra plataforma no mundo que possua a quantidade e a qualidade dos *corpora*, dos recursos e extensões disponíveis no software, nem a organização estrutural e infraestrutural pensada para o estudo do léxico e para o trabalho do lexicógrafo, mais especificamente. Toda a disposição do website é direcionada para a pesquisa prática, automatizada e totalmente digital, dispensando a necessidade de manter anotações manuais ou arquivos salvos no computador, pois o espaço de armazenamento disponível na plataforma é suficiente para atividades lexicográficas, sendo possível, inclusive, manter diversos *corpora* autorais armazenados e criar links e páginas para dicionários elaborados no site.

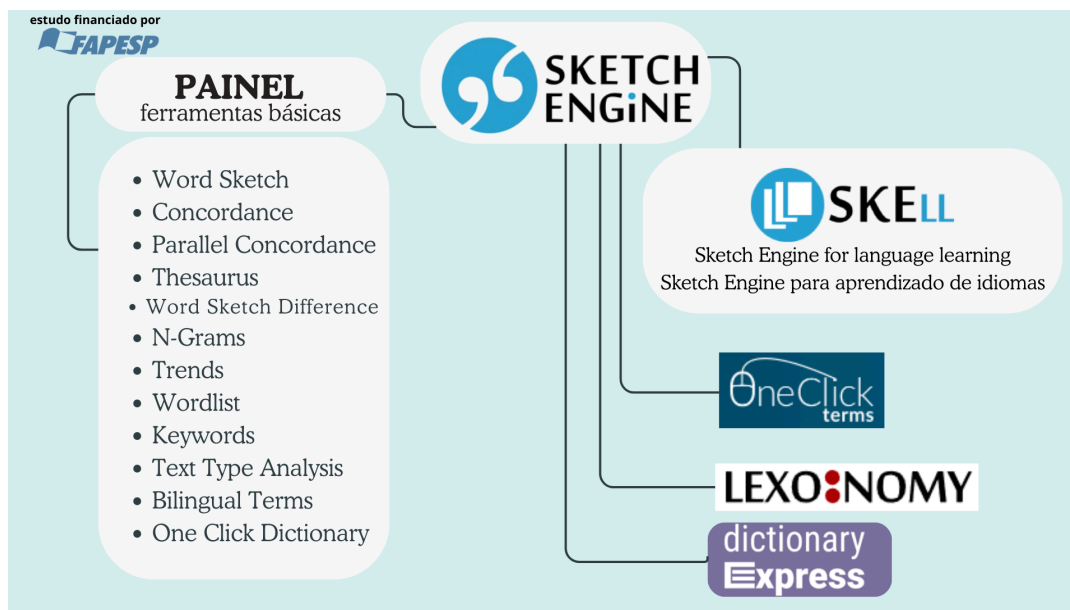
Mediante as informações apresentadas, busca-se evidenciar a profícua união dos avanços tecnológicos à prática lexicográfica. Cada vez mais, estudos e pesquisas nesse campo do saber devem inserir na metodologia teórica e prática o que há de mais avançado e disponível. A pesquisa e a divulgação de recursos como o SE são capazes de

potencializar, reorganizar e estabelecer novos critérios de criação para projetos lexicográficos, tais quais os de elaboração de dicionários digitais ou o uso de ferramentas como o *SKELL* para fornecer recursos para estudantes de língua inglesa, como se pretende averiguar e mapear durante o estudo.

2.1 Ferramentas e extensões disponíveis na plataforma Sketch Engine

A SKE possui duas vertentes principais de recursos, a saber, as ferramentas disponíveis na página inicial do website (*Dashboard*) — *Word Sketch*, *Thesaurus*, *Parallel Concordance*, *N-Grams*, *Trends*, *OneClick Dictionary*, *Word Sketch Difference*, *Concordance*, *Wordlist*, *Keywords*, *Text Type Analysis* e *Bilingual Terms* — e as extensões — *SKELL*, *One Click Terms*, *Dictionary Express* e *Lexonomy*. As extensões encontram-se em websites independentes, mas cujas funcionalidades correlacionam-se com as ferramentas principais encontradas no painel da página inicial da SKE. O quadro explicativo a seguir demonstra a distribuição de cada recurso mencionado:

Quadro 1. Sketch Engine e ramificações



Fonte: Elaborado pelas autoras.

Passa-se então a descrever cada um dos recursos, iniciando-se pelas ferramentas e seguindo para as extensões.

2.1.1 Ferramentas do painel inicial do website

É essencial esclarecer que para o uso de qualquer ferramenta da SKE é necessário selecionar o *corpus* na barra superior da página inicial. O tipo de *corpus* determinará as opções de ferramentas disponíveis. Os *corpora OpenSubtitles*, por exemplo, possuem seus correspondentes em outros idiomas; a escolha de um deles dará a opção de uso da ferramenta *Parallel Concordance*. Ao escolher o *corpus*, as ferramentas passíveis de uso serão destacadas automaticamente na página inicial. A depender da escolha do *corpus*, as opções de ferramentas podem ser:

- a) *Word Sketch* é a função que dá nome à plataforma, pode ser traduzido como “Raio X da palavra”: trata-se de gerar dados que expõem o comportamento gramatical e combinatório de determinada palavra (lema)⁷. A partir da palavra ou combinatória, a ferramenta explora o *corpus* selecionado em busca de padrões recorrentes de seu comportamento léxico-sintático. O resultado é apresentado separado por classe de palavras ou função sintática.
- b) *Concordance* (concordanciador) é uma ferramenta que desempenha uma função básica na maior parte das plataformas de análise de *corpus*, que serve para gerar listas de sentenças com a palavra ou palavras de busca. A pesquisa de concordância na SKE é feita no mesmo modelo que uma pesquisa no buscador Google, ou seja, os usuários inserem uma palavra ou palavras simples e a ferramenta interpreta e faz a busca no *corpus* selecionado. A ferramenta é insensível ao uso de maiúsculas e minúsculas na palavra (se o usuário pesquisar pelo termo ‘rio’, a ferramenta encontrará ‘Rio’ e ‘RIO’, por exemplo); e pode encontrar tanto o lema (ou seja, a forma-base da palavra), quanto as suas respectivas derivações e variações (se o usuário pesquisar o lema ‘fazer’ verbo, encontrará também suas formas, como ‘faço’, ‘fez’, ‘feito’, ‘faria’ e demais variações). O concordanciador também é capaz de buscar combinatórias de palavras separadas por espaço, como ‘fazer hora’ ou ‘fazer a barba’.
- c) *Parallel Concordance* (Concordanciador em paralelo) desempenha a mesma função do concordanciador, porém, busca e apresenta dados que permitem a comparação tradutória em mais de 30 idiomas simultaneamente. A ferramenta gera listas em paralelo das mesmas sentenças contendo a palavra ou palavras de busca e dois ou mais idiomas. É possível buscar por palavras, frases, *tags*, documentos, metadados, estruturas de *corpus* em *corpora* de mesmo texto em diferentes idiomas. A busca tende a ser mais lenta quanto maior o *input* da busca e maior o número de idiomas, mas ainda assim, é oferecida uma dinamicidade que supera uma pesquisa manual tradicional. A funcionalidade somente funciona com *corpora* especializado para a função, os nomeados como *parallel*. Existem as opções básica e avançada. A pesquisa simples de palavra ou combinatória de palavras se dá por meio de um *input* no

⁷ ‘lema’, no âmbito dos estudos lexicais, refere-se à palavra em sua forma base, como no caso de verbos, a forma infinitiva. Observa-se que na SKE, às vezes, como é o caso das buscas na Word Sketch, a palavra lema é usada como sinônimo de ‘palavra’.

idioma-base do *corpus* selecionado, na barra “search”. Em seguida, é possível selecionar os demais idiomas nos quais deseja-se visualizar a tradução. É possível verificar traduções específicas por meio da opção *translated as (optional)*, na qual pode-se sugerir a tradução desejada, e da opção “in”, na qual seleciona-se o idioma disponível. Existe um botão com símbolo de adição logo em seguida, no qual se pode acrescentar outros idiomas progressivamente. Na pesquisa avançada, é possível definir o idioma e o tipo de entrada, o que permite encontrar estruturas mais específicas no *corpus*. Utilizando a opção CQL (*Corpus Query Language*), o usuário pode buscar resultados complexos de estruturas gramaticais e lexicais. Um exemplo do uso da ferramenta em seu modo simples no *corpus* “OpenSubtitles 2018 parallel - Brazilian Portuguese” é a de procurar pela palavra “ansioso” na barra de pesquisa e de sua correspondente tradução para o inglês “excited”. Esta função é útil para a construção de glossários com temática específica e também para dicionários bilíngues, sobretudo os pedagógicos.

- d) *Thesaurus* (Tesouro) desempenha função semelhante a um dicionário de sinônimos e agrupa palavras de uma mesma categoria semântica. O Tesouro da SKE pode agrupar as palavras em forma de lista e em nuvens de palavras, de uma forma distribucional. O software observa as relações lexicais entre diferentes grupos de palavras e sua frequência, sendo capaz de listar e fornecer qual a ocorrência e a frequência de um certo termo no *corpus*, encontrando as correspondências mais próximas e adequadas, possíveis substituições e sinônimos ou palavras com proximidade semântica.
- e) *Word Sketch Difference* é semelhante ao *Word Sketch*. Trata-se de uma ferramenta elaborada para gerar e revelar comportamentos de duas ou mais palavras favorecendo a análise comparativa entre elas. A busca pode ocorrer por meio de três opções disponíveis: ‘*lemma*’, que compara dois lemas diferentes através de suas combinatórias, ‘*word forms*’, que observa as colocações de duas formas diferentes do mesmo lema, e ‘*subcorpora*’, que compara o uso de um mesmo lema dentro de dois *subcorpora* do mesmo *corpus* também através de combinatórias. Por meio dessas informações, é possível compreender como a palavra se comporta, como é registrada e utilizada e a qual campo semântico pertence, obtendo clareza a respeito de seus conceitos, acepções e usos. Os dados são apresentados separadamente utilizando cores diferentes para cada palavra ou *subcorpus*, permitindo observar qual combinação é a mais típica (em verde) ou a menos usual (em cores neutras para colocações sem preferência ou vermelha para usos raros ou incomuns). Essa ferramenta também pode fornecer uma lista de frequência com dados quantitativos.
- f) *N-Grams* é a ferramenta específica de processamento de texto para obter listas de frequência e sequências de *tokens*⁸, de bigramas, trigramas ou superiores, encontrando “blocos” de significado, como expressões, o que faz a ferramenta ser também conhecida como ‘*multiword expressions*’ (expressões multilexicais). O usuário pode filtrar pelas opções com os atributos que deseja exibir ou ocultar e o software é capaz de realizar a ‘tokenização’, a ‘lematização’ e a classificação (‘*tagging*’) do *corpus* de

⁸ Um *token* é a menor unidade que compõe um *corpus*, pode ser uma palavra ou uma pontuação, um dígito, uma abreviação, etc.

forma automática e com rapidez no idioma fornecido, além de inúmeros outros filtros para listagem de *tokens*, palavras, expressões e mais. O sistema da SKE é feito para gerar as listas com rapidez, em alguns segundos. O processamento pode ser um pouco mais lento a depender da complexidade das expressões ou do tamanho do *corpus* (quando esse conter mais de um bilhão de palavras, por exemplo).

- g) *Trends* (Tendências) é uma ferramenta que permite realizar análises sincrônicas e diacrônicas de palavras e pode ser utilizada para observar neologismos, padrões de uso de uma língua através do tempo, auxiliando na identificação de variações e de mudanças linguísticas. Por esta razão, só funciona com *corpora* que possuem conteúdo que permite análises temporais, os *corpora* categorizados como “*Trends*”, tais como o “*English Trends Corpus*” ou o “*Portuguese Trends*” — esses *corpora* são elaborados extraindo conteúdos de fontes dinâmicas, como notícias, *feeds* ou redes sociais que o software obtém por meio de *crawlers* (robôs que fazem esta varredura na *web*) e armazenamento de dados.
- h) *Wordlist* (listas de palavras), como o nome sugere, realiza a listagem de palavras e indica as suas frequências no *corpus*, permitindo buscas por classes gramaticais, palavras ou lemas, ou ainda, letras que aparecem no início, no final ou em qualquer parte da palavra na lista. A ferramenta pode gerar uma lista que combina classes gramaticais, uso de outros critérios lexicais em uma mesma lista. Atuando no nível de *tokens*, ela faz a separação entre palavras e não-palavras, excluindo as não-palavras automaticamente, o que a diferencia dos *N-Grams*. A lista pode exibir apenas as palavras e sua frequência ou também o lema e os “*tags*” (abreviações que fornecem informações morfológicas sobre a palavra).
- i) *Keywords* (palavras-chave) é um extrator que permite identificar dados únicos ou típicos num certo *corpus*. A ferramenta pode ser usada para extrair terminologias de um certo campo semântico, traduções ou interpretações; pode extrair uma palavra ou uma unidade plurilexical que são típicas de um *corpus* ou documento, para estabelecer seus tópicos, conteúdos e assunto; e pode, ainda, comparar dois *corpora* distintos, para pontuar o que há de original e único em um *corpus* em foco A quando comparado com um *corpus* de referência B. Os resultados são apresentados em forma de palavras-chave e grupos de palavras, com *hiperlinks* que direcionam às respectivas sentenças nos *corpora*. Pode ser especialmente útil no trabalho terminológico ou na tradução de documentos, por exemplo.
- j) *Text type analysis* é a ferramenta que serve para mostrar metadados, ou seja, informações descritivas a respeito de um *corpus*. Ela oferece transparência, mostrando quais e quantos documentos, palavras e *tokens* compõem os textos selecionados para compor os *corpora*, dependendo das informações disponíveis neles. Essa ferramenta é composta por filtros que exibem as estatísticas em formato de um gráfico de setores. Por meio dessa “análise de tipo textual”, é possível extrair os metadados do *corpus* e as estruturas nas quais estão contidos, o número de estruturas ou *tokens* que o *corpus* comporta, dados estatísticos gerais ou de um *subcorpus* específico e links para verificar concordância. Também é possível fazer *downloads* e filtrar resultados.

- k) *Bilingual Terms* é a ferramenta responsável por identificar automaticamente pares de termos equivalentes em duas línguas com base em *corpora* paralelos, ou seja, aqueles elaborados a partir de textos traduzidos alinhados frase a frase, o que permitirá analisar a ocorrência simultânea de palavras e expressões em segmentos correspondentes de cada língua. A partir desses dados, pode-se avaliar candidatos a termos bilíngues e organizar listas terminológicas com estatísticas a respeito do grau de associação entre eles, providenciando a identificação de equivalentes terminológicos com base em evidências empíricas, utilizando dados estatísticos para detectar padrões reais de correspondência em textos paralelos, o que garante confiabilidade para os resultados sugeridos, que serão validados por humanos para verificação final.
- l) *One Click Dictionary* (Dicionário a um clique): Embora conste dentre as ferramentas no painel inicial, o ícone é um atalho para a extensão *Lexonomy* e serve para criar dicionários. É um software para a escrita de dicionário por meio do *Lexonomy*. Para utilizá-lo, é necessário criar uma conta no *Lexonomy* e uma chave API, que serve para identificar e autorizar o acesso ao software.

2.1.2 Extensões

Além das ferramentas do painel inicial, a SKE desenvolveu quatro extensões para finalidades específicas, que são assim denominadas, pois possuem websites próprios, embora continuem sendo ramificações da tecnologia e das contas de usuário da SKE. As extensões disponibilizadas são SKELL, One Click Terms, Lexonomy e Dictionary Express, abaixo detalhadas quanto às funcionalidades que desempenham.

- a) *SKELL*, sigla para ‘*Sketch Engine for Language Learners*’ (*Sketch Engine* para aprendizes de idiomas), é um website que funciona como uma versão facilitada da SKE, com a interface configurada como uma barra de pesquisa simples. Conta com as ferramentas *Examples* (‘Exemplos’, similar à ferramenta de Concordância); *Collocations* (versão simplificada de *Word Sketch*, capaz de fornecer combinatórias e listas de palavras) e *Synonyms* (Sinônimos, que é o Tesouro cujos dados são extraídos do acervo de *corpora* SKE). Seu design é pensado para ser acessível e intuitivo, fornecendo as informações com organização visual e explicações em linguagem clara, com menor uso de termos lexicográficos. Até a data desta pesquisa, a SKELL estava disponível nos idiomas Inglês, Tcheco, Alemão, Estoniano, Italiano e Russo. Não necessita de login para ser utilizada.
- b) *One Click Terms* é um extrator automático de terminologia disponível em diversos idiomas para consultas em *corpora* especializados. Sua função principal consiste em identificar termos em determinado *corpus* (num conjunto de textos de determinada área técnica ou científica, por exemplo). A extensão utiliza métricas estatísticas e técnicas de extração terminológica para determinar as combinações de palavras que possuem com maior relevância em *corpus*, obtendo padrões linguísticos e mostrando os trechos de onde os extraiu no *corpus*. Também encontra equivalentes bilíngues quando aplicada a *corpora* paralelos. Dessa forma, pode pré-elaborar um rascunho da

terminologia com informações editáveis, como classificação gramatical, equivalentes lexicais e possíveis significados. Seu público-alvo é composto por tradutores técnicos, pesquisadores, terminólogos e outros profissionais de áreas análogas, pois proporciona a identificação de termos num banco de dados de forma rápida e baseada em evidências. A extensão serve para descobrir e organizar termos relevantes para alimentar projetos lexicográficos, classificando-se como uma ferramenta analítica. É importante esclarecer que inicialmente, tratava-se de um recurso online na página inicial da SKE (o *One Click Dictionary*), mas foi atualizado e adquiriu um domínio de website próprio, passível de ser acessado por meio de login na conta da SKE ou de forma independente.

- c) *Lexonomy* é uma plataforma *online* e gratuita elaborada especificamente para a escrita de produtos lexicográficos como e-dicionários, glossários, listas de terminologia e outros. Trata-se de um ambiente digital preparado para redigir verbetes e que permite criar e organizar um dicionário digital em um mesmo ambiente virtual. A *Lexonomy* destaca-se pela flexibilidade, pois permite ao usuário criar seus próprios verbetes e personalizá-los, criando estruturas para a inserção de dados, compatíveis com o tipo de dicionário em elaboração (bilíngue, pedagógico, terminológico etc.). Sua estrutura possibilita a exportação dos dados da SKE e também promove o trabalho colaborativo, permitindo que os usuários compartilhem os dados e editem em conjunto. A estrutura para a redação e revisão de verbetes é intuitiva e clara, o que dispensa profundo conhecimento em programação. Não serve para gerar dados de forma automática, apenas armazena e estrutura os dados nele alimentados, mas os *corpora* podem ser analisados previamente na SKE. A integração entre a *Lexonomy* e a SKE é proveitosa e recomendada, uma vez que otimiza o trabalho do lexicógrafo, poupando tempo e eventuais gastos com bancos de dados e outras opções digitais pagas para a criação e armazenamento de produtos lexicográficos.
- d) *Dictionary Express* é a extensão direcionada para a publicação e acesso a e-dicionários previamente armazenados e disponibilizados no meio digital. Provê hospedagem e visualização para que bases lexicográficas armazenadas se tornem dicionários *online* pesquisáveis, com link próprio, de maneira a ser acessível para o usuário final. Por isso, destaca-se por permitir que instituições e pesquisadores lancem seus e-dicionários sem que precisem criar um site a partir do zero, otimizando o tempo dedicado ao trabalho lexicográfico. Esta extensão é definida pela equipe da *Lexical Computing* como um método de edição e publicação de dicionários, o qual combina processamento de linguagem natural, linguística de *corpus*, inteligência artificial e validação de dados manuais para a elaboração de trabalhos lexicográficos. Por meio da extensão, podem ser produzidos dicionários, glossários, bases de dados terminológicos, listas de palavras, listas de sinônimos, dentre outros. A principal vantagem proporcionada pelo método é a otimização de tempo e custos de produção de conteúdo, além da variedade de possibilidades de publicação, em conteúdos da web, leitores digitais (*e-readers*), aplicativos diversos, serviços e pesquisas de tradução e interpretação de idiomas e até em jogos de palavras. De modo simplificado, a aplicação de todas as demais ferramentas e ramificações da SKE culminará em produtos passíveis de serem divulgados pelo *Dictionary Express*.

A descrição das ferramentas e extensões revela, por fim, que a SKE foi desenvolvida para apoiar o trabalho lexicográfico desde seu projeto até a sua publicação e decorrentes atualizações.

2.2 Plataforma Sketch Engine — questões práticas

Ilustrar-se-á, na sequência, o uso prático da SKE, demonstrando a aplicação da funcionalidade da ferramenta *Parallel Concordance* recomendada para estudos linguísticos bilíngues. Neste estudo, propõe-se a busca de equivalente para o verbo 'ir' com vistas à redação de um verbete na direção português-inglês. Usou-se a ferramenta *Parallel Concordance* para gerar sentenças correlatas em português e inglês que permitiram analisar o padrão tradutório do verbo.

Ao buscar as sentenças em paralelo que contenham o verbo 'ir' e correspondentes em inglês, foi possível notar que, de 45 sentenças geradas, 31 trouxeram o equivalente *go*, enquanto as demais trouxeram *leave* (4 sentenças), *going* (2), *get* (2), *come* (2), *head for* (1), *let* (2) e *follow* (1). A análise motivou a decisão lexicográfica de incluir como equivalentes de 'ir' no dicionário, os verbos *go*, *leave* e *head*, sendo que, para esses dois últimos, impõe-se a decisão de incluir uma nota explicativa do contexto de uso e, no caso de *leave*, um comentário de que em inglês é comum enfatizar o ato de deixar o lugar e não o de ir a um outro – se alguém está em casa e vai para o supermercado, por exemplo, e isso já é de conhecimento das pessoas à sua volta, no português é comum despedir-se dizendo: — Estou indo! Enquanto em inglês é comum despedir-se dizendo: — *I am leaving!* (Estou saindo.). A ficha inicial do verbete foi redigida da seguinte maneira:

Tabela 1. Ficha inicial do verbete "ir"

Lema: Ir
Equivalentes: Go / Leave / Head

Fonte: Elaborado pelas autoras.

A funcionalidade *Parallel Concordance* proporcionou observar o comportamento e os padrões tradutórios do verbo 'ir'. Para complementar a redação do verbete, usou-se novamente a ferramenta *Parallel Concordance* e a ferramenta secundária GDEX.

Realizou-se a mesma busca por sentenças em paralelo que continham o verbo ‘ir’ e correspondentes em inglês. Ao clicar no ícone GDEX, na parte superior da página com os resultados gerados, o sistema gerou nova lista com sentenças em português e respectivas sentenças equivalentes em inglês recomendadas como candidatas a exemplos lexicográficos. A partir da listagem, foram selecionadas aquelas que melhor atendessem à função e ao público do dicionário. No caso, considerou-se como critério de inclusão de exemplos: (1) sentenças simples ou coordenadas; (2) sentenças com vocabulário de níveis A1 e A2 em inglês, (assim consideradas no *Oxford Learner's Dictionary*, onde cada palavra possui indicação de nível CEFR); (3) sentenças nas estruturas afirmativas, negativas e interrogativas sempre que possíveis.

Para o verbete ‘Ir’, foi gerada a listagem (ver anexos) e foram selecionados os seguintes exemplos:

Tabela 2. Listagem de exemplos para o verbo “ir”.

IR- GO

Você pode ir para a escola de chinelos - *You can go to school with slippers.*

Não podemos ir mais rápido - *We can't go any faster.*

Podemos ir para outro lugar? - *Can we go somewhere else?*

Fonte: Elaborado pelas autoras.

Ao incluí-los na ficha do verbete tem-se:

Tabela 3. Ficha do verbete ‘ir’ com exemplos.

Lema: IR

Equivalentes: *Go / Leave / Head*

Exemplos:

Afirmativa: Você pode ir para a escola de chinelos. - *You can go to school with slippers.*

Negativa: Não podemos ir mais rápido - *We can't go any faster.*

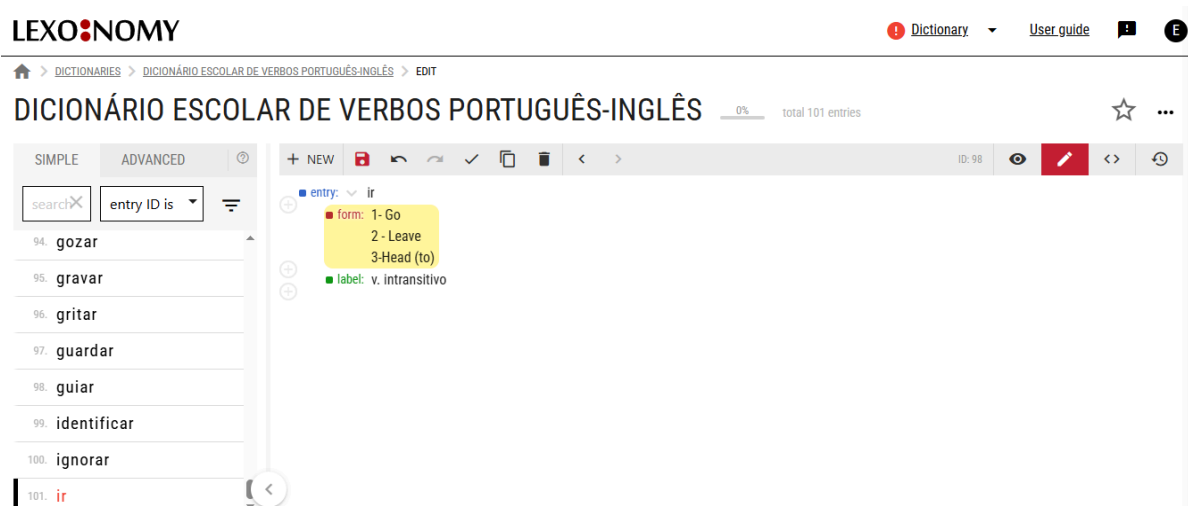
Interrogativa: Podemos ir para outro lugar? - *Can we go somewhere else?*

Fonte: Elaborado pelas autoras.

Por meio dos procedimentos apresentados para a seleção dos equivalentes e redação do verbete IR, foi possível demonstrar como a SKE pode ser utilizada para gerar dados para análise e fundamentação de obra lexicográfica.

Além de contribuir para a escrita dos verbetes, a SKE permite elaborar e publicar um dicionário na extensão *Lexonomy*. Para tanto, primeiramente, é informada a nomenclatura (macroestrutura) e, posteriormente, passa-se para a edição verbete a verbete (microestrutura). O processo de inclusão de verbetes e respectivas informações pode ser feito a longo prazo e, em conjunto, por membros de uma mesma equipe de lexicógrafos. A Figura 1 demonstra a elaboração do verbete ‘IR’.

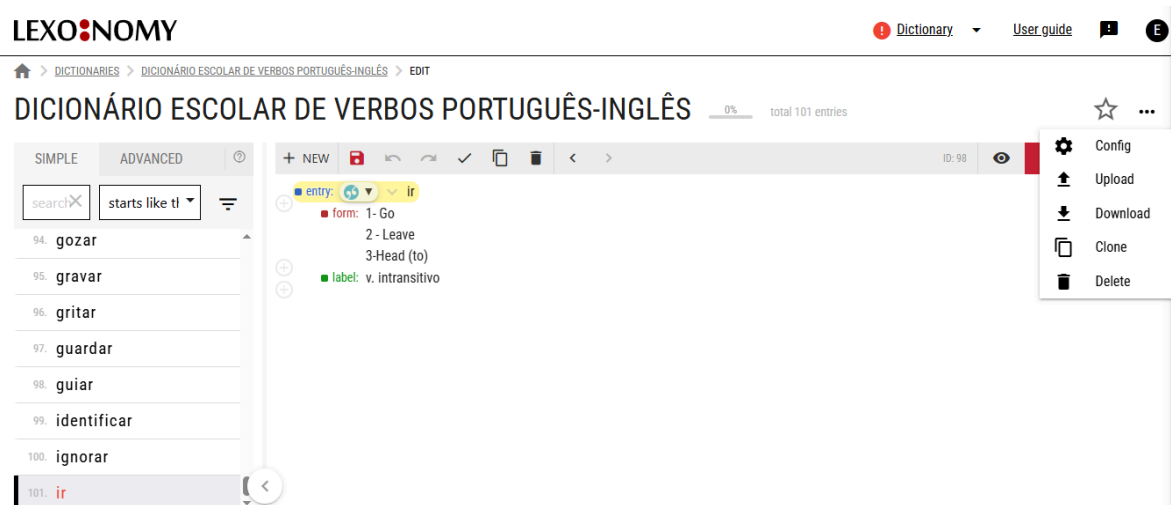
Figura 1. Verbetes verbo ‘ir’



Fonte: Rascunho do dicionário no *Lexonomy*, conta das autoras. Acesso em dez. 2025.

A estrutura do verbete pode ser ajustada para atender ao projeto da obra e há a possibilidade de se inserir links e valer-se de outras opções da extensão. A Figura 2 revela as opções de configuração da estrutura do verbete.

Figura 2. Opções de configuração do verbete



Fonte: Rascunho do exemplo de dicionário, *Lexonomy*, conta das autoras. Acesso em dez. 2025.

Após a redação dos verbetes, o dicionário poderá ser publicado e disponibilizado para uso sem custos para o seu armazenamento e atualizações na própria plataforma *Lexonomy*.

A presente pesquisa concluiu que a plataforma é indiscutivelmente eficiente na elaboração de dicionários e de produtos lexicográficos em geral, porque é a única plataforma destinada a fins lexicográficos (*Lexicographical computing*). Além disso,

01. disponibiliza *corpora* prontos para uso;
02. disponibiliza *corpora* em mais de 100 idiomas;
03. apresenta uma gama de opções de ferramentas primárias, secundárias e extensões que se correlacionam e automatizam a busca, garantindo a acuracidade de resultados;
04. vincula-se a extensões que permitem a redação, a editoração, a publicação e a atualização de obras lexicográficas;
05. disponibiliza tutoriais e vídeos que asseguram a eficácia de seu uso e contribuem para o letramento lexicográfico digital;
06. disponibiliza contato direto e suporte atencioso da equipe técnica.

Acredita-se que o estudo colabora para a demanda pela edificação de conhecimento teórico e prático sobre lexicografia computacional em nosso país. O

conhecimento dos avanços e-lexicográficos é o único meio para garantir que obras lexicográficas de base computacional sejam registradas e amplamente viabilizadas no universo linguístico digital.

Conclusão

Conclui-se que a SKE contribui com excelência para o trabalho lexicográfico, assegurando confiabilidade na geração de dados que fundamentam as pesquisas e análises lexicográficas. A perspicácia humana segue imprescindível para o estudo da linguagem, pois será ela que identificará lacunas a serem investigadas e questões que promovam a evolução e a construção do conhecimento.

Além do mais, saber, conhecer e explorar os *corpora* linguísticos existentes nos idiomas contemporâneos promove competências técnicas necessárias ao ambiente digital e auxilia na compreensão e no avanço da e-lexicografia.

Revelam-se boas prospecções a respeito do futuro da e-lexicografia e do trabalho lexicográfico por meio da integração de novos recursos tecnológicos. A SKE mostra-se um instrumento sem igual, versátil e acessível para o desenvolvimento de atividades lexicográficas.

Referências

ANTCONC. Disponível em: <https://www.laurenceanthony.net/software/antconc/>. Acesso em: 06 jul. 2025.

CLARIN ERIC. Disponível em: <https://www.clarin.eu/>. Acesso em: 20 jul. 2025.

DEEPL PRO. Disponível em: <https://www.deepl.com/en/pro>. Acesso em: 05 jul. 2025.

DICTIONARY EXPRESS. Disponível em: <https://dictionary.express/>. Acesso em: 11 jul.2025.

GRANGER, S. *Electronic lexicography: from challenge to opportunity*. In: GRANGER, S.; PAQUOT, M. (org.). **Electronic Lexicography**. Oxford: Oxford University Press, p. 1-11, 2012.

HÖFLING, C.; ZACARIAS, R. A. S. O uso de inglês como meio de instrução: impacto na internacionalização, mobilidade acadêmica e formação de estudantes. In: LUCAS, O. P.; RODRIGUES, R. F. L. (org.). **Temas e Rumos nas Pesquisas em Linguística (Aplicada):** Questões empíricas, éticas e práticas. Campinas: Pontes, v. 2, p. 255-279, 2017.

KILGARRIFF, A., BAISA, V., BUŠTA, J., JAKUBÍČEK, M., KOVÁŘ, V., MICHELFEIT, J., RYCHLÝ, P., SUCHOMEL, V. *The sketch engine: Ten years on*. **Lexicography: Journal of ASIALEX**, 1(1), 7–36, 2014.

KILGARRIFF, A.; KOSEM, I. *Corpus tools for lexicographers*. In S. G. M. Paquot (Ed.), **Electronic lexicography**. Oxford, Oxford University Press, p. 31–56, 2012.

KILGARRIFF, A.; RYCHLÝ, P.; SMŘŽ, P.; TUGWELL, D. *The Sketch Engine*. **Proceedings of the 11th EURALEX International Congress**, p. 105–116, 2004.

LANDAU, S. I. **Dictionaries: the art and craft of lexicography**. 2. ed. Cambridge: Cambridge University Press, 2001.

LEW, R.; SCHRYVER, G.-M. de. *Dictionary users in the digital revolution*. **International Journal of Lexicography**, v. 27, n. 4, p. 341–359, 2014.

LEXONOMY. Disponível em: <https://www.lexonomy.eu/#home>. Acesso em: 11 jul. 2025.

ONE CLICK TERMS. Disponível em: <https://terms.sketchengine.eu/>. Acesso em: 12 jul. 2025.

SKELL - Sketch Engine for Language Learners. Disponível em: <https://skell.sketchengine.eu/#home?lang=en>. Acesso em: 11 jul. 2025.

SKETCH ENGINE. Disponível em: <http://www.sketchengine.eu>. Acesso em: 11 jul. 2025.

TAGNIN, S. E. O.; ZACARIAS, R. A. S. (org.). **O alcance da linguística de corpus: do léxico ao discurso**. XV Encontro de Linguística de Corpus. Livro eletrônico. São Paulo: Ed. dos Autores, 2023.

VIANA, V. **Verbos modais em contraste: análise de corpus da escrita de universitários em inglês**. 2008. 230 f. Dissertação (Mestrado em Letras) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

VIANA, V.; TAGNIN, S. E. O. **Corpora no ensino de línguas estrangeiras**. São Paulo: Hub, 2011.

WORDSMITH TOOLS. Disponível em: <https://www.lexically.net/wordsmith/>. Acesso em: 15 jul. 2025.

Anexo



simple ir • 6,325,160

8,980.37 per million tokens • 0.9%



①	<s> Então, aqui vai o que aconteceu. </s>	<s> Yep, okay, so here's what happened. </s>
①	<s> E uma dessas coisas foi que eu preciso mudar. </s>	<s> And, um, one of the things I realized is that I need to break my old patterns. </s>
①	<s> - Entendível, e eu vou ter que passar por isso. </s>	<s> Understandable, and I will just have to transcend that. </s>
①	<s> Então eu vou começar com esses lixos. </s>	<s> So, um, that starts with, uh, trashing these. </s>
①	<s> Acho que eu deveria ir . </s>	<s> Well, I guess I should go. </s>
①	<s> Eu poderia ir pra casa e beber perfume. </s>	<s> I could go home and drink some mouthwash. </s>
①	<s> Vamos conversar. </s>	<s> Let's-let-let's talk. </s>
①	<s> Então eu vou embora. </s>	<s> So I'm gonna go. </s>
①	<s> Vamos fazer isso. </s>	<s> Let's do this. </s>
①	<s> Você foi na loja de etiquetas? </s>	<s> Did you go to a bumper sticker store? </s>
①	<s> - vou guardar. </s>	<s> - so I will keep this. </s>
①	<s> Mas o resto vai vai embora, adeus. </s>	<s> But the rest of this is going, bye. </s>
①	<s> Mas o resto vai vai embora, adeus. </s>	<s> But the rest of this is going, bye. </s>
①	<s> Foi bem fácil, ao menos pra mim. </s>	<s> Was pretty easy, at least for me. </s>
①	<s> Você foi aceito na Harvard do Sul, não é? </s>	<s> You got into the Harvard of the South, after all, which you somehow managed to mention on your test. </s>
①	<s> Você foi aceito em Emory, Greg. </s>	<s> You got into Emory, Greg. </s>
①	<s> Claro, um dia eu irei </s>	<s> ♪ Which I will eventually, of course ♪ </s>
①	<s> Isso não vai te levar a lugar algum </s>	<s> ♪ It'll get you nowhere in life ♪ </s>
①	<s> - Pode atender quem for ... </s>	<s> - Can you get whoever that... </s>
①	<s> - Sim, vou lá. </s>	<s> - Yeah, I will get that. </s>
①	<s> Vamos conversar depois, está bem? </s>	<s> And... just... </s><s> let's continue the convo later, okay? </s>
①	<s> E se for os Testemunhas de Jeová, diga a eles que eu não creio e nunca vou crer. </s>	<s> And if it's the people across the street from that church, just tell them I don't believe and will never believe. </s>
①	<s> E se for os Testemunhas de Jeová, diga a eles que eu não creio e nunca vou crer. </s>	<s> And if it's the people across the street from that church, just tell them I don't believe and will never believe. </s>
①	<s> Não foi nada. </s>	<s> Oh, it was no problem at all. </s>
①	<s> Infelizmente a carta era melhor do que o teste dela, mas ela foi à Cal State Northridge, que é a Harvard do norte. </s>	<s> Unfortunately the letter was better than her application, but she did get into Cal State Northridge, which is the Harvard of Northridge. </s>
①	<s> Por que não vamos ...? </s>	<s> Oh. </s><s> Okay, hey, why don't we just... ? </s>
①	<s> Isso iria matar a Valencia. </s>	<s> Okay, this would kill Valencia. </s>
①	<s> Vamos , Brody. </s>	<s> Come on, Brody. </s>
①	<s> Vamos ver os vestidos? </s>	<s> We ready to try on some dresses? </s>
①	<s> Jas, vamos . </s>	<s> Jas, come on. </s>
①	<s> Vamos ver os vestidos. </s>	<s> Let's go pick out the first round of dresses. </s>
①	<s> Valencia, oi, isso vai ser tão divertido. </s>	<s> Hey, Valencia, hi, oh, uh, this is going to be so fun. </s>
①	<s> Vou deixar as coisas bem claras aqui. </s>	<s> I want to make myself very, very clear. </s>
①	<s> Josh contou que foi você. </s>	<s> Josh told me it was all you. </s>
①	<s> No final de tudo isso, eles vão querer te adotar. </s>	<s> Yeah, yeah, like, by the end of this, they're going to want to adopt you. </s>
①	<s> Ele com certeza iria cobrir os meus peitões. </s>	<s> Yeah, those would definitely cover most of my nipples. </s>
①	<s> Eu vou me render e dizer que isso é gracioso. </s>	<s> I'm gonna go out on a limb and say that's, like, really gracious. </s>

i	<s> Vocês se divertam com os seus corpos magros e saudáveis, enquanto eu vou achar uma capa de botijão de gás. </s>	<s> So you kids can go crazy with your halters and your sheaths and your cutouts, and I will wear a dress fit to upholster a chair. </s>
i	<s> Como foi na loja de noivas com a Valencia? </s>	<s> Okay, so how did it go with Valencia at the dress shop? </s>
i	<s> Minha ideia foi colocar pó de mico no vestido dela. </s>	<s> My first thought was poison oak in the dress. </s>
i	<s> - Não vamos fazer isso. </s>	<s> We're not doing any of that. </s>
i	<s> - Não vamos . </s>	<s> - We're not. </s>
i	<s> Eu vou ser só mais uma moça que foi convidada de última hora. </s>	<s> I'm going to be like any other teenage ex-boyfriend's sister's last-minute bridesmaid. </s>
i	<s> Eu vou ser só mais uma moça que foi convidada de última hora. </s>	<s> I'm going to be like any other teenage ex-boyfriend's sister's last-minute bridesmaid. </s>
i	<s> Vamos voltar ao caso. </s>	<s> Sorry. </s><s> Let's get back to the case. </s>
i	<s> Quando irá pegar os vestidos? </s>	<s> When are you getting the bridesmaid gowns? </s>
i	<s> Porque conheço um cara no Maine que vai me mandar o pó de mico. </s>	<s> 'Cause I got a guy in Maine who's going to send me some primo itchy, sticky plants. </s>
i	<s> E ele disse que vai mandar um casal de carrapatos por mais bufunfa... </s>	<s> And he said he was going to throw in a couple deer ticks for, you know, a little extra... </s>
i	<s> Eu digo vamos ver o que pode acontecer, talvez alguma doença. </s>	<s> - Oh. </s><s> - I say we roll the dice for, you know, some possible Lyme disease. </s>
i	<s> Tenho que ir . </s>	<s> I have to go. </s>

Fonte: *Sketch Engine*, conta das autoras. Acesso em dez. 2025.

Sobre as autoras

Eloá Bera Santana

Orcid: <https://orcid.org/0009-0008-3340-610X>

Mestranda em Estudos Linguísticos pela Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), IBILCE, de São José do Rio Preto, no Programa de Pós-Graduação em Estudos Linguísticos, na linha de Lexicologia e Lexicografia (2026). Licenciada em Letras com dupla habilitação (Português/Inglês) pela UNESP - Faculdade de Ciências e Letras de Assis (2025). Iniciação Científica com apoio FAPESP em Lexicografia, na área de Linguística (2025)

Regiani Aparecida Santos Zacarias

Orcid: <https://orcid.org/0000-0001-5212-3895>

Professora Livre-Docente do Departamento de Letras Modernas da UNESP/Assis e docente do Programa de Mestrado Profissional da UNESP/Bauru. Doutora em Estudos da Linguagem (UEL, 2011), com Pós-Doutorado em Estudos Linguísticos pela UNESP/Araraquara. Pesquisa Metalexigrafia e Lexicografia Bilíngue Pedagógica, com apoio FAPESP e CNPq. Lidera o Grupo de Pesquisa CNPq Tradução, Léxico e Interculturalidade. Fulbright Scholar-in-Residence no Santa Fe College, EUA (2012-2013).