Filol. Linguíst. Port., São Paulo, v. 21, n. 2, p. 271-288, jul./dez. 2019
http://dx.doi.org/10.11606/issn.2176-9419.v21i2p271-288

Linguística de corpus: teoria, perspetivas metodológicas e ensino das línguas

Corpus linguistics: theory, methodological perspectives and language teaching

Carlos Assunção[*]
Universidade de Trás-os-Montes, Vila Real, Portugal

Carla Araújo[**]
Instituto Politécnico de Bragança, Bragança, Portugal

Resumo: A palavra 'corpus' reveste-se de ambiguidade, uma vez que diz respeito, por um lado, a um conjunto de dados e, por outro, a um conjunto de métodos. Em relação ao sentido de conjunto de dados, verifica-se que todos os linguistas são potenciais utilizadores de *corpora*, já que a Linguística é uma disciplina empírica. Os métodos sobre *corpora*, construídos a partir de simples concordâncias, são especialmente estatísticos e/ou probabilísticos. O contexto contemporâneo de ensino-aprendizagem das línguas impõe o desenvolvimento e a disponibilização de recursos didáticos, como as bases de dados que sirvam de suporte às exigentes e rigorosas práticas pedagógicas da conjuntura educativa do século XXI. Simultaneamente, este recurso permite ainda ao professor obter as competências necessárias para fomentar junto dos alunos uma atitude crítica e reflexiva sobre a língua, tendo em vista o desenvolvimento da capacidade de observação e análise da língua num processo de descoberta do seu sistema de funcionamento.

Palavras-chave: Linguística de Corpus. Metodologias. Recursos didáticos. Ensino.

Abstract: The word 'corpus' is ambiguous, since it involves, on the one hand, a set of data and, on the other hand, a set of methods. In relation to the sense of dataset, all linguists are potential users of *corpora*, since linguistics is an empirical discipline. The methods of research on *corpora*, constructed from simple concordances, are especially of a statistical and / or probabilistic nature. The contemporary context of teaching and learning of languages imposes the development and availability of didactic resources, such as databases that support the demanding and rigorous pedagogical practice of the educational context of the 21st century. Simultaneously, this resource also allows the teacher to obtain the necessary skills to foster a critical and reflexive attitude towards the language among the students, bearing in mind the development of their ability to observe and analyze the language in a process of a discovery of its working system.

Keywords: Corpus linguistics. Methodologies. Teaching resources. Teaching.

[*] Professor Catedrático, Investigador Integrado do Centro de Estudos em Letras. O CEL é uma unidade de investigação financiada pela Fundação para a Ciência e a Tecnologia (UID/LIN/00707/2020); cassunca@utad.pt

[**] Professora Adjunta, Investigadora Integrada no Centro de Estudos em Letras. O CEL é uma unidade de investigação financiada pela Fundação para a Ciência e a Tecnologia (UID/LIN/00707/2020); carla.araujo@ipb.pt

1 INTRODUÇÃO

A Linguística de Corpus ancora-se num paradigma teórico que se caracteriza por uma abordagem empirista e por uma conceção da linguagem como um sistema probabilístico. Em Linguística, o empirismo é uma abordagem que concede estatuto primordial aos dados que provêm da observação da linguagem, geralmente agrupados sob a forma de corpus, opondo-se ao racionalismo, que se baseia no estudo da linguagem a partir da introspeção, entendida como maneira de averiguar modelos de funcionamento estrutural e a formação do processo cognitivo da linguagem. Relativamente ao sentido de um conjunto de métodos, podemos dizer que a pesquisa sobre corpus implica mais métodos indutivos do que métodos hipotético-dedutivos, por conseguinte as análises conduzidas pelos dados (*data-driven*) são preferidas em detrimento das análises conduzidas por regras (*rule-driven*).

Este tema já foi alvo de diversos estudos. Investigadores como Stubbs (1993), Sinclair (2004), Leech (1992), Kennedy (1998, 2005), Tognini-Bonelli (2001), McEnery e Wilson (1996), Halliday (2006), Chomsky (1956; 1964), Teubert (2005), Meyer (2002), Bowker e Pearson (2002), McEnery et al. (2006), Sardinha (2004), Gries (2010), entre outros, levaram a cabo estudos no âmbito da Linguística de Corpus e do ensino das línguas, como a seguir veremos.

Na sequência destas reflexões, faremos o levantamento das posturas filosóficas características da conceção empirista e racionalista da linguagem, representadas pelos seus maiores vultos. Por um lado, Halliday (2006), representante da conceção empirista, e, por outro lado, Chomsky (1964), o maior vulto do racionalismo na Linguística. Constataremos também que outros linguistas, como, por exemplo, Kennedy (1998), apresentam-se a favor de uma abordagem mista, aliando intuição e corpus. Além disso, em certas vertentes, também se aproximam de Chomsky, aceitando que o funcionamento da linguagem não pode ser revelado na sua plenitude pelos *corpora*, já que os mesmos não possibilitam a distinção entre as estruturas possíveis e as estruturas impossíveis. Indo ao encontro de muitos dos linguistas de corpus, Kennedy admite que o não surgimento de um certo elemento num corpus, ainda que seja de grande extensão, não invalida a sua existência. De modo contrário, o aparecimento de uma estrutura num corpus não determina, de modo automático, a sua gramaticalidade.

Problematizar-se-á se a Linguística de Corpus é uma metodologia ou uma teoria ou se fica entre ambas, seguindo as perspetivas dos melhores teorizadores, a utilização de recursos computacionais na linguística de corpus e o ensino das línguas, ainda no plano teórico mas já com alusão a alguns recursos disponíveis, e finalizaremos com as concordâncias e sua utilização na sala de aula. Neste último tópico utilizaremos alguma da melhor bibliografia disponível e indicaremos algumas das vantagens das concordâncias suportadas por estudos sobre o assunto. A metodologia utilizada neste trabalho consiste na revisão bibliográfica dos estudos mais reconhecidos internacionalmente sobre esta área da Linguística de Corpus com o objetivo de que esta reflexão possa servir como suporte para um melhor entendimento dos tópicos acima enunciados e que possam problematizar esta área do conhecimento de forma a que, posteriormente, se façam análises a partir destes recursos que a Linguística de Corpus proporciona.

2 LINGUÍSTICA DE CORPUS: ENTRE A TEORIA E METODOLOGIA

A dificuldade que envolve a definição de Linguística de Corpus como uma teoria ou como uma metodologia tem sido debatida a partir de diferentes pontos de vista. Tem sido argumentado que a Linguística de Corpus não é verdadeiramente um domínio de pesquisa, mas apenas uma base metodológica para estudar a linguagem. No entanto, muitos linguistas que trabalham com corpus tendem a concordar que a Linguística de Corpus vai muito para além desse papel exclusivamente metodológico. Por exemplo, Tognini-Bonelli (2001, p. 1) situa a Linguística de Corpus na esfera da Linguística Aplicada e concede-lhe um estatuto teórico, referindo algumas áreas que beneficiam da Linguística de Corpus:

> In this context we take the view that although corpus linguistics belongs to the sphere of applied linguistics, it differs from other partner disciplines under the same umbrella in that it can be seen as a pre-application methodology. [...] Corpus linguistics has, therefore, a theoretical status and because of this it is in a position to contribute specifically to other applications. Among the areas which have benefited from the input of corpus linguistics are lexicography, language teaching, translation, stylistics, grammar, gender studies, forensic linguistics, computational linguistics, to quote but a few[1].

Relativamente à controvérsia em relação à Linguística de Corpus e considerando-a como um "exercício altamente teórico", Halliday (2006, p. 130) refere o seguinte:

> At a recent conference devoted to modern developments in corpus studies, I was struck by the way that a number of speakers at the conference were setting up an opposition between "corpus linguists" and "theoretical linguists" - not a conflict, I mean, but a distinction, as if these were members of two distinct species. I commented on this at the time, saying that I found it strange because corpus linguistics seemed to me to be, potentially at least, a highly theoretical pursuit. Work based on corpus studies has already begun to modify our thinking about lexis, about patterns in the vocabulary of languages; and it is now beginning to impact on our ideas about grammar. In my view, this impact is likely to be entirely beneficial. Corpus linguistics brings a powerful new resource into our theoretical investigations of language[2].

[1] Neste contexto, consideram que, embora a linguística de corpus pertença à esfera da linguística aplicada, ela difere de outras disciplinas parceiras sob o mesmo âmbito, na medida em que pode ser vista como uma metodologia de pré-aplicação. [...] A linguística de corpus tem, portanto, um estatuto teórico e, por isso, ela está em condições de contribuir especificamente para outras aplicações. Entre as áreas que beneficiaram do contributo da linguística de corpus estão a lexicografia, o ensino de línguas, a tradução, a estilística, a gramática, os estudos de género, a linguística forense, a linguística computacional, para citar apenas algumas (tradução nossa).

[2] Numa conferência recente, dedicada à moderna evolução em estudos de corpus, fiquei impressionado com a maneira como uma série de oradores da conferência fizeram uma oposição entre "linguistas de corpus" e "linguistas teóricos" - não um conflito, eu quero dizer, mas uma distinção, como se fossem membros de duas áreas distintas. Eu comentei sobre isso no momento, dizendo que achei estranho, porque a linguística de corpus me pareceu ser, pelo menos potencialmente, um exercício altamente teórico. O trabalho com base em estudos de corpus já começou a modificar o nosso pensamento sobre o léxico, sobre os padrões no vocabulário das línguas; e agora está a começar a ter impacto sobre as nossas ideias sobre gramática. Na minha opinião, esse impacto é suscetível de ser inteiramente

McEnery e Wilson (1996, p. 2) respondem de forma afirmativa e negativa à seguinte questão: A Linguística de Corpus será um ramo da Linguística?

> The answer to this question is both yes and no. Corpus linguistics is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics, and so on. All of these disciplines concentrate on describing/explaining some aspect of language use. Corpus linguistics in contrast is a methodology rather than an aspect of language requiring explanation or description. [...] Corpus linguistics is a methodology that may be used in almost any area of linguistics, but it does not truly delimit an area of linguistics itself. Corpus linguistics does, however, allow us to differentiate between approaches taken on the study of language, and in that respect it does define an area of linguistics, or at least a series of areas of linguistics. Hence we have corpus-based syntax as opposed to non-corpus-based-syntax, corpus-based semantics as opposed to non-corpus-based-semantics, and so on. So, while corpus linguistics is not an area of linguistic enquiry in itself, it does, at least, allow us to discriminate between methodological approaches taken to the same area of enquiry by different groups, individuals or studies[3].

Esta resposta tem sido fornecida por diversos autores que consideram a Linguística de Corpus como uma abordagem e uma metodologia no campo de ação das diversas áreas da Linguística.

Gries (2010) afirma que a relação entre a linguística de corpus e a teoria linguística tem sido um pouco problemática, apontando duas razões para essa situação: "corpus linguists differ as to what they think CL is: a tool, method(ology), discipline, theory, paradigm, framework, …; there are some things that make CL appear less attractive to the observer from theoretical linguistics"[4] (2010, p. 41).

Teubert (2005, p. 5) considera que a "corpus linguistics looks at phenomena which cannot be explained by recourse to general rules and assumptions"[5]. Gries (2010, p. 331) refere que "many corpus linguists who are interested in explaining

benéfico. A linguística de corpus traz um novo recurso, poderoso, para as nossas investigações teóricas da linguagem (tradução nossa).

[3] A resposta a esta pergunta é sim e não. A linguística de corpus não é um ramo da linguística, no mesmo sentido da sintaxe, da semântica, da sociolinguística, e assim por diante. Todas estas disciplinas se concentram em descrever/explicar algum aspeto do uso da língua. A linguística de corpus, em contraste, é uma metodologia, em vez de um aspeto da linguagem que exige explicação ou descrição. [...] A linguística de corpus é uma metodologia que pode ser usada em praticamente qualquer área da linguística, mas não verdadeiramente delimitar uma área da própria linguística. No entanto, a linguística de corpus permite-nos diferenciar entre as abordagens adotadas no estudo da linguagem, e, nesse aspeto, faz definir uma área da linguística, ou, pelo menos, uma série de áreas da linguística. Assim, temos sintaxe baseada em corpus, em oposição à sintaxe não baseada em corpus, semântica baseada em corpus, em oposição à semântica não baseada em corpus, e assim por diante. Portanto, embora a linguística de corpus não seja uma área de pesquisa linguística em si mesma, ela permite, pelo menos, discriminar entre abordagens metodológicas adotadas na mesma área de pesquisa por diferentes grupos, indivíduos ou estudos (tradução nossa).

[4] Os linguistas de corpus divergem quanto ao que consideram que a linguística de corpus é: uma ferramenta, um(a) método(ologia), uma disciplina, uma teoria, um paradigma, um enquadramento, ...; existem algumas coisas que fazem a linguística de corpus parecer menos atraente para o observador da linguística teórica (tradução nossa).

[5] A linguística de corpus olha para fenómenos que não podem ser explicados pelo recurso a regras e a pressupostos gerais (tradução nossa).

phenomena this way, especially since 'general rules and assumptions' does not rule out probabilistic rules and assumptions"[6].

No âmbito dos autores que consideram a Linguística de Corpus uma teoria, destacam-se, por exemplo, Leech (1992, p. 106), Stubbs (1993, p. 2), Tognini-Bonelli (2001, p. 1) e Teubert (2005, p. 2). Este último refere que a Linguística de Corpus não é em si um método:

> Corpus linguistics is not in itself a method: many different methods are used in processing and analysing corpus data. It is rather an insistence on working only with real language data taken from the discourse in a principled way and compiled into a corpus. However, one should be wary of using such data merely to find out more about what we know already, since what (we think) we know is often derived from pre-corpus study. Corpus data provide insights of a type which has not previously been available. Concepts and categories derived from introspective language study or from models taken from other fields (e.g. computation) may not be appropriate for describing real language data (Teubert, 2005, p. 4)[7].

Remetendo a Linguística de Corpus para as "contingências do uso da língua", Teubert salienta que a mesma olha para fenómenos que não podem ser explicados pelo recurso às regras e pressupostos gerais:

> Corpus linguistics does not have its starting point in language universals if we understand universals as ontological features (and not as theoretical concepts). Little is reliably known about the language faculty all human beings share. The study of this language faculty is outside the remit of corpus linguistics. Rather, corpus linguistics looks at phenomena which cannot be explained by recourse to general rules and assumptions. It is primarily concerned with the contingencies of language use. Normally, we become aware of language only if there is a communication disorder. These disorders have their origins in the variation we find within and between discourses. They can be analysed in terms of the differences we observe between one language use and another (Teubert, 2005, p. 5)[8].

[6] Muitos linguistas de corpus estão interessados na explicação de fenómenos desta forma, especialmente porque 'regras e premissas gerais' não excluem regras e premissas probabilísticas (tradução nossa).

[7] A linguística de corpus não é, em si, um método: muitos métodos diferentes são usados no processamento e análise de dados de corpus. É, antes, uma insistência em trabalhar apenas com dados reais da linguagem, retirados do discurso, de uma forma baseada em princípios, e compilados num corpus. No entanto, deve-se ter cuidado, ao usar esses dados apenas para descobrir mais sobre o que já sabemos, já que o que (pensamos que) sabemos é frequentemente derivado de estudos pré-corpus. Os dados do corpus fornecem tipos de informações que não estavam disponíveis anteriormente. Os conceitos e as categorias derivados do estudo introspetivo da linguagem ou de modelos obtidos de outros campos (por exemplo, computação) podem não ser apropriados para descrever os dados da linguagem real (tradução nossa).

[8] A linguística de corpus não tem o seu ponto de partida nos universais da linguagem, se entendermos universais como características ontológicas (e não como conceitos teóricos). Pouco se sabe sobre a faculdade da linguagem que todos os seres humanos compartilham. O estudo desta faculdade da linguagem está fora do âmbito da linguística de corpus. Em vez disso, a linguística de corpus olha para fenómenos que não podem ser explicados pelo recurso a regras e a pressupostos gerais. Está principalmente preocupada com as contingências do uso da linguagem. Normalmente, tomamos conhecimento da linguagem somente se houver um distúrbio de comunicação. Esses distúrbios têm as suas origens na variação, que encontramos dentro e entre discursos. Eles podem ser analisados em termos das diferenças que observamos entre um uso linguístico e outro uso (tradução nossa).

Segundo Teubert (2005), uma vez que a Linguística não é uma ciência como as ciências naturais, cuja missão é a procura da "verdade", não há verdadeiro significado.

> Linguistics is not a science like the natural sciences whose remit is the search for 'truth'. It belongs to the humanities, and as such it is a part of the endeavour to make sense of the human condition. Interpretation, and not verification, is the proper response to the quest for meaning. There is no true meaning (Teubert, 2005, p. 7-8)[9].

Por conseguinte, o linguista de corpus não faz juízos de valor sobre o que é permitido e o que não é permitido:

> The corpus linguist is not privileged as an 'expert' to pass judgment on what is permissible and what not. He or she is part of a discourse community, not outside of it. Corpus linguists have to submit their findings to their discourse community and argue for their acceptance. The discourse community is, in principle, a democratic community. Every member has the right to contribute to the discourse, and to discuss, modify or reject what other members say. The discourse organises itself. All regimentation from the outside strangles the creativity of the discourse Community (Teubert, 2005, p. 8)[10].

Para Teubert (2005, p. 13), a Linguística de Corpus não é uma teoria, mas uma "metodologia imperfeita":

> Corpus linguistics is and will remain an imperfect methodology to make sense of the discourse. For me, it is not so much a theory of language as a conceptual frame for studying the transmission of content in a discourse community, as evidenced in the intertextuality of the discourse. Corpus linguistics localises the study of language, once again, firmly and deliberately, in the *Geisteswissenschaften*, the humanities[11].

Teubert (2010, p. 24), relativamente ao tipo de ciência em que se insere a Linguística de Corpus, refere que John Sinclair, o decano da Linguística de Corpus, pouco fez para se estabelecer como um membro proeminente da comunidade *linguística teórica*, considerando que o mesmo não estava realmente interessado na questão de saber em que tipo de ciência a Linguística de Corpus se insere.

---

[9] A linguística não é uma ciência como as ciências naturais, cuja missão é a busca da 'verdade'. Pertence às humanidades e, como tal, faz parte do esforço de entender a condição humana. Interpretação, e não verificação, é a resposta adequada à questão do significado. Não existe um verdadeiro significado (tradução nossa).

[10] O linguista de corpus não tem o privilégio de ser um 'especialista' para julgar o que é permitido e o que não é. Ele ou ela fazem parte de uma comunidade discursiva, não estão fora dela. Os linguistas de corpus precisam de enviar as suas descobertas para a comunidade do discurso e argumentar pela sua aceitação. A comunidade do discurso é, em princípio, uma comunidade democrática. Todo o membro tem o direito de contribuir com o discurso e discutir, modificar ou rejeitar o que os outros membros dizem. O discurso organiza-se. Todo o regime de fora estrangula a criatividade do discurso da Comunidade (tradução nossa).

[11] A linguística de corpus é, e continuará a ser, a metodologia imperfeita de fazer o sentido do discurso. Para mim, não é tanto uma teoria da linguagem como um quadro conceptual para estudar a transmissão de conteúdo numa comunidade discursiva, como evidenciado na intertextualidade do discurso. A linguística de corpus localiza o estudo da linguagem, mais uma vez, com firmeza e deliberadamente, nas humanidades, as ciências humanas (tradução nossa).

> John Sinclair, the doyen of corpus linguistics, did little to establish himself as a prominent member of the theoretical linguistics community. He did not look for confrontation there. As far as I know, he was not really interested in the question whether corpus linguistics belongs somehow to the 'hard' sciences or the interpretive sciences, what the French and Germans call the sciences humaines or Geisteswissenschaften (Teubert, 2010, p. 24)[12].

Segundo Teubert (2010, p. 24), para Sinclair, a Linguística de Corpus constituiu uma maneira nova e diferente de olhar para a linguagem: "For [Sinclair], it was a new and different way to look at language"[13].

Em relação aos autores que consideram a Linguística de Corpus como uma metodologia, referimos, por exemplo, McEnery e Wilson (1996), Meyer (2002), Bowker e Pearson (2002), McEnery et al. (2006). Para os autores anteriormente referidos, a Linguística de Corpus, embora seja um sistema de princípios e métodos, não é uma teoria em si mesma "[…] corpus linguistics is a whole system of methods and principles of how to apply *corpora* in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory in itself" (McEnery et al., 2006, p. 7)[14].

No entanto, os que reivindicam para a Linguística de Corpus o estatuto de paradigma teórico em si mesmo, relativamente às abordagens chomskianas ou a muitas escolas de Linguística Cognitiva, são uma minoria:

> It quickly became clear that those of us who insisted that corpus linguistics is a theoretical paradigm in its own right, incompatible with other mainstream approaches such as Chomskyan mentalism or the many schools of cognitive linguistics, are in a minority (Teubert, 2010, p. 19)[15].

Teubert afirma que a Linguística de Corpus é encarada como uma metodologia, e não como uma teoria, não só pelos linguistas cognitivos, mas também por aqueles que têm como objetivo o PNL, processamento de linguagem natural:

> It is obvious that those advocating corpus linguistics as a methodology, not a theory, do not all call themselves cognitive linguists. But they all would argue that they are working within a scientific framework on par with that of the natural sciences. This seems to be the case even if they regard themselves less as scientists and more as engineers, their aim being NLP, natural language processing, i. e. developing language technology applications such as machine translation or artificial intelligence. Language

[12] John Sinclair, o decano da linguística de corpus, fez pouco para se estabelecer como um membro proeminente da comunidade linguística teórica. Ele não procurou confronto, lá. Até onde eu sei, ele não estava realmente interessado na questão de saber se a linguística de corpus pertence, de alguma forma, às ciências 'duras' ou às ciências interpretativas, o que os franceses e alemães chamam de ciências humanas ou humanidades (tradução nossa).

[13] Para [Sinclair], era uma maneira nova e diferente de ver a linguagem (tradução nossa).

[14] A linguística de corpus é um sistema completo de métodos e princípios de como aplicar *corpora* nos estudos da linguagem e no ensino/aprendizagem, certamente tem um status teórico. No entanto, o status teórico não é uma teoria em si (tradução nossa).

[15] Logo ficou claro que aqueles de nós que insistiam que a linguística de corpus é um paradigma teórico por si só, incompatível com outras abordagens convencionais, como o mentalismo chomskyano ou as muitas escolas de linguística cognitiva, são uma minoria (tradução nossa).

> engineering, like any other kind of engineering, is commonly understood as the application of scientific results for practical purposes. Indeed we can observe throughout the 20th century a persistently reiterated 'mantra' that the study of language or languages has its place in the 'hard' sciences. But what exactly are the 'scientific' results, the undeniable 'facts' about language that theoretical linguistics, the linguistics of the language system, have taught us? How does corpus linguistics compare to this invocation of 'hard' science? (Teubert, 2010, p. 21)[16].

Para Teubert (2010, p. 23), a linguística que pretende ser reconhecida como uma "*hard science*" focaliza o sistema da linguagem, não o discurso.

Para exemplificar as diferenças entre a Linguística de Corpus e a Linguística Aplicada, Teubert (2010, p. 23) estabelece uma analogia com a meteorologia:

> […] as meteorology is more than collecting weather data; it is about modelling a system that can predict what is happening next by applying a rule-based mechanism to the data. Applied meteorology is about forecasting; theoretical meteorology is about modelling the mechanism in ever finer detail that makes forecasts possible. Applied linguistics, on the other hand, is not about what the next text will be about; it is about telling people what they have to do if they want to avoid breaking rules, if they want to be understood[17].

Retomando a analogia com a meteorologia, Teubert (2010, p. 28) apresenta as diferentes perspetivas do linguista cognitivo e do linguista de corpus:

> Cognitive linguists, of course, have a different perspective. What a corpus linguist looks at is only the surface, they would claim, is comparable to the data of temperature, pressure, wind or precipitation a meteorologist would record. Forecasting needs such data, but it needs more than that: There must be a theoretical model, a system behind it that tells the meteorologist to predict how the weather is going to develop on the basis of such data. Likewise, for language, a cognitivist would claim, there must be some mechanism, partly innate and partly acquired, that enables people to say

[16]É óbvio que aqueles que defendem a linguística de corpus como metodologia, não como teoria, nem todos se chamam linguistas cognitivos. Mas, todos eles argumentam que estão a trabalhar dentro de uma estrutura científica parecida com a das ciências naturais. Parece ser esse o caso, mesmo que se considerem menos cientistas e mais engenheiros, com o objetivo de PNL, processamento de linguagem natural, isto é, desenvolvendo tecnologia de aplicativos de linguagem, como a tradução automática ou a inteligência artificial. A engenharia da linguagem, como qualquer outro tipo de engenharia, é comumente entendida como a aplicação de resultados científicos para fins práticos. De facto, podemos observar, ao longo do século XX, um 'mantra', persistentemente reiterado, de que o estudo da linguagem, ou das línguas, tem o seu lugar nas ciências 'duras'. Mas, quais são exatamente os resultados 'científicos', os inegáveis 'factos' sobre a linguagem que a linguística teórica, a linguística do sistema de línguas, nos ensinou? Como a linguística de corpus se compara a essa invocação da ciência 'dura'? (tradução nossa).

[17][C]omo a meteorologia, é mais do que coletar dados meteorológicos; trata-se de modelar um sistema que pode prever o que vai acontecer a seguir, aplicando um mecanismo baseado em regras aos dados. A meteorologia aplicada é sobre previsão; a meteorologia teórica trata da modelagem do mecanismo em detalhes cada vez mais refinados, que possibilitam previsões. A linguística aplicada, por outro lado, não é sobre o que será o próximo texto; trata-se de dizer às pessoas o que elas devem fazer, se quiserem evitar violar as regras, se quiserem ser entendidas (tradução nossa).

> certain things but not others. Only if we can obtain a better grasp of this mechanism, we have begun to understand the idea of language[18].

Tal como a meteorologia é mais do que a recolha de dados sobre os estados do tempo, também a Linguística de Corpus é mais do que a recolha de dados linguísticos. Deste modo, o linguista de corpus pretende encontrar estruturas e padrões nos dados, no discurso, no corpus. Por outro lado, os linguistas teóricos principais acreditam que esses dados nunca serão suficientes para sabermos por que motivo certas coisas podem ser ditas e outras não. Os linguistas cognitivos encontram as suas respostas na mente. Valorizando o discurso, Teubert (2010, p. 33) assume o seu problema com a linguística cognitiva:

> The problem I have with cognitive linguistics is that it provides an interpretation of what I would like to call introspection. As so much research into folk psychology has shown, introspection can very easily be deceptive (cf. Christensen and Turner, 1993). What we find looking into our minds is only too often what the discourse has taught us to expect. Of course there is also psycholinguistic experimentation. In the seventies of last century, such experiments 'proved' that the processing speed of the mind was strictly correlated to the number of transformations needed to turn a Chomskyan deep structure into a surface structure. It took years to defeat these claims[19].

3 UTILIZAÇÃO DE RECURSOS COMPUTACIONAIS NA LINGUÍSTICA DE CORPUS E O ENSINO DAS LÍNGUAS

No âmbito dos recursos computacionais utilizados na Linguística de Corpus, Mello e Sousa (2012, p. 5-6) referem os seguintes: frequenciadores, programas para listar palavras, que procedem à contagem das palavras do corpus e fornecem listas de frequência de formas. As formas individuais são conhecidas como tipos - *types* e as suas ocorrências, como *tokens*; concordanciadores, programas que possibilitam a procura de palavras particulares num corpus, fornecendo listas para as ocorrências das mesmas em contexto; etiquetadores, que realizam a análise automática do corpus e inserem etiquetas (códigos) de cariz morfossintático, sintático, semântico, prosódico ou discursivo; ferramentas de engenharia textual, pacotes de software que procuram modularizar as diversas atividades de processamento de linguagem natural.

[18]Os linguistas cognitivos, é claro, têm uma perspetiva diferente. O que um linguista de corpus olha é apenas a superfície, afirmariam eles, é comparável aos dados da temperatura, da pressão, do vento ou da precipitação, que um meteorologista registaria. A previsão precisa desses dados, mas precisa mais do que isso: deve haver um modelo teórico, um sistema por trás dele que diga ao meteorologista para prever como o tempo se vai desenvolver, com base nesses dados. Da mesma forma, para a linguagem, reivindicaria um cognitivista, deve haver algum mecanismo, parcialmente inato e parcialmente adquirido, que permita às pessoas dizer certas coisas, mas não outras. Somente se pudermos compreender melhor esse mecanismo, começamos a entender a ideia de linguagem (tradução nossa).

[19]O problema que tenho com a linguística cognitiva é que ela fornece uma interpretação do que eu gostaria de chamar de introspeção. Como muitas pesquisas em psicologia popular demonstraram, a introspeção pode facilmente enganar (cf. Christensen e Turner 1993). O que descobrimos, olhando em nossas mentes, é, com muita frequência, o que o discurso nos ensinou a esperar. Claro que também há experimentação psicolinguística. Nos anos setenta do século passado, tais experimentos 'provaram' que a velocidade de processamento da mente estava estritamente correlacionada com o número de transformações necessárias para transformar uma estrutura profunda de Chomsky em uma superfície. Levou anos para derrotar essas alegações (tradução nossa).

São muito diversificados os recursos disponíveis que permitem a exploração de *corpora*. Há programas que permitem quer a compilação de *corpora*, quer o tratamento de *corpora*, como, por exemplo, o TextSTAT, que permite compilar *corpora* buscando textos na *web* ou em pastas específicas e fornece listas de formas e a sua frequência e respetivas concordâncias[20].

A compilação de *corpora* através da *web* pode também ser realizada por meio do *Bootcat*[21]. O Bootcat possui *scripts* que permitem buscar páginas específicas na *web*, partindo de palavras-chave. A utilidade do Bootcat é ilimitada tanto para a compilação de *corpora* especializados, como para corpora paralelos para tradução ou *corpora* para fins lexicográficos, entre outros.

Outro exemplo de recurso disponível que permite a exploração de *corpora* é o concordanceador *AntConc*[22]. O *AntConc* dispõe de um conjunto de ferramentas que permitem listar as linhas de concordância de uma dada forma, mostrar a linha de concordância em contexto, visibilizar o arquivo de texto, listar *clusters*, listar frequências, palavras-chave e colocados.

No âmbito das ferramentas disponíveis que permitem a exploração de *corpora* em português, podemos utilizar, de forma gratuita, por exemplo, o VISL[23].

Para proceder à compilação e à anotação, podem seguir-se parâmetros distintos. No entanto, atualmente, pretende-se a adoção de diretrizes facilitadoras da padronização dos critérios adotados. Essas diretrizes[24] pretendem padronizar a representação de textos em formato digital, por exemplo, nos documentos do *Text Encoding Initiative*.

O *Expert Advisory Group on Language Engineering Standards* – EAGLES[25] constitui também um exemplo do objetivo de padronização dos critérios relacionados com o tratamento de *corpora*.

O corpus tem que ser muito bem adaptado ao objetivo do ensino com o apoio da tecnologia, conduzindo a uma aprendizagem mediada, que se caracteriza pela construção do conhecimento aluno/computador, ultrapassando a visão tradicional da transmissão de conhecimento professor/aluno. O trabalho com *corpora* na sala de aula acarreta uma aproximação entre as práticas de investigação e as práticas de ensino-aprendizagem. O aluno adquire o papel de um investigador que pretende obter respostas a partir dos dados disponíveis no corpus.

[20] Encontra-se disponível em http://neon.niederlandistik.fu-berlin.de/en/textstat/.

[21] Ver em http://bootcat.sslmit.unibo.it/).

[22] Encontra-se no sítio http://www.antlab.sci.waseda.ac.jp/software.html.

[23] Ver em http://beta.visl.sdu.dk/visl/pt/.

[24] Para uma visão mais detalhada, veja-se http://www.tei-c.org/index.xml.

[25] Para uma visão mais detalhada, veja-se http://www.ilc.cnr.it/EAGLES/browse.html.

Apesar de todas as vantagens do uso de *corpora* na sala de aula, Sardinha (2004) refere que a influência da Linguística de Corpus, no ensino, se verifica apenas de modo indireto:

> A influência da Linguística de Corpus no ensino ocorre de modo indireto, apenas quando os resultados da pesquisa são absorvidos e incorporados, em geral parcialmente, nos materiais de ensino. A entrada no ambiente pedagógico ocorre, primordialmente, pelos livros didáticos e muito pouco por intermédio do professor (2004, p. 255).

O autor recomenda que os alunos sejam expostos a exemplos fornecidos por falantes nativos da língua em estudo, a fim de poderem ultrapassar dificuldades linguísticas que não estão previstas pelas abordagens pensadas com propósitos pedagógicos. Neste sentido, o léxico extraído de situações linguísticas autênticas adquire primordial importância, devido às possíveis colocações e combinações distintas das de natureza artificial, que foram criadas com objetivos pedagógicos.

Leech (1997) também releva a utilidade do acesso a *corpora* por parte dos alunos, tendo em vista a exploração dos mesmos partindo dos objetivos de pesquisa dos próprios. No entanto, as tarefas de acesso e análise de *corpora* devem ser guiadas, maioritariamente, pelo professor, que desempenha a função de orientador e facilitador do processo. Atendendo ao facto de o acesso do aluno ao corpus envolver determinadas etapas e certas limitações, o professor deve promover uma progressão gradual no uso de *corpora* na sala de aula, para que a aprendizagem da língua por descoberta possa acontecer.

De acordo com Sardinha (2004, p. 254-255), o impacto gerado pela acessibilidade e pela exploração de *corpora* de línguas naturais no ensino é suscetível de ser resumido em quatro grandes áreas de atuação: "a descrição da linguagem nativa; descrição da linguagem do aprendiz; transposição de metodologias de pesquisa académica para a sala de aula; desenvolvimento de materiais de ensino, currículos e abordagens".

Concedendo uma relevância às concordâncias e considerando também outras vertentes, a área da *transposição de metodologias de pesquisa académica para a sala de aula* encontra-se associada à área do *desenvolvimento de materiais de ensino, currículos e abordagens*. No campo de ação da área que diz respeito "à criação de metodologias de ensino, inspiradas na exploração de *corpora* ou em conceitos da Linguística de Corpus" (Sardinha, 2004, p. 255), salientamos as seguintes três abordagens lexicais: o *Lexical Syllabus* (Currículo Lexical), o *Lexical Approach* (Abordagem Lexical) e o *DDL – Data Driven Learning* (Ensino Movido a Dados), que apresentaremos seguidamente.

O *Lexical Syllabus*, criado por Dave Willis (1990), constitui uma abordagem que se fundamenta na perspetiva de um corpus composto por textos produzidos por falantes nativos da língua em estudo, preconizando que os sentidos mais usuais da língua são manifestados por meio do vocabulário mais frequente.

Do ponto de vista pedagógico, um dos benefícios desta abordagem deve-se ao facto de os alunos se identificarem com o conteúdo estudado, de um modo mais rápido. Neste sentido, não é só o corpus a orientar a produção de materiais, mas também as próprias abordagens de descoberta que são incorporadas como princípios

didáticos. Por isso, a gramática não se apresenta como uma fonte válida fornecedora de dados que orientem a criação de um curso de línguas.

O *Lexical Syllabus* possui como princípos norteadores os seguintes: criar materiais baseados em corpus; ensinar o que é mais frequente; conceber léxico e gramática como um todo; usar linguagem autêntica. De acordo com Sardinha (2004, p. 282-286), o material produzido para criar livros didáticos baseou-se no mesmo corpus utilizado para realizar o primeiro dicionário *Cobuild*, com 7,3 milhões de palavras.

O *Lexical Approach* foi concebido por Michael Lewis (2000) e caracteriza-se pelo desenvolvimento de atividades em que o léxico assume também centralidade. A proposta de Lewis (2000, p. 132) baseia-se em colocações:

> Collocations […] co-occur naturally, and the first task of the language teacher is to ensure that they are not unnecessarily taken apart in the classroom. If words occur together, learners need to notice that co-occurrence and, if they are to be recorded in a vocabulary book, the words should be recorded together[26].

Os alunos, através da identificação de colocações em textos, procedem ao registo das novas palavras ou *chunks*[27] em cadernos didáticos. Segundo Lewis (2000), uma vez que determinados itens lexicais se combinam com mais frequência com outros itens particulares, constituindo combinações frequentes, é fundamental permitir aos alunos o contacto com essas colocações, não os impedindo de aceder aos contextos autênticos de ocorrência de tais itens.

Em relação à abordagem *DDL – Data Driven Learning,* termo criado por Tim Johns (1991) para denominar o ensino, nomeadamente de línguas estrangeiras, o aluno tem acesso a dados linguísticos extraídos de linguagem autêntica (concordâncias), com o objetivo de os analisar, colocar hipóteses, fazer inferências e generalizações, construindo uma aprendizagem por descoberta. Tim Johns desenvolveu este tipo de abordagem com o objetivo de ultrapassar as dificuldades manifestadas pelos alunos estrangeiros de pós-graduação da universidade de Birmingham, destacando o léxico e a gramática através de concordâncias, chamando a atenção para colocações e padrões léxico-gramaticais. Esta abordagem constitui, por outro lado, um estímulo ao desenvolvimento da autonomia dos discentes e da aprendizagem por descoberta.

A abordagem *DDL – Data Driven Learning* caracteriza-se por utilizar linhas de concordância para analisar a língua; centra-se nas relações e nos padrões léxico-gramaticais; fornece ao aluno o papel de descobridor/investigador; e concede ao professor a função de proporcionar ao aluno os meios que lhe permitam realizar descobertas a partir da observação das concordâncias. Deste modo, o aluno descobre

[26]As colocações [...] coocorrem naturalmente e a primeira tarefa do professor de língua é garantir que não sejam, desnecessariamente, desmontadas na sala de aula. Se as palavras ocorrem juntas, os alunos precisam de observar essa coocorrência e, se elas devem ser registadas num livro de vocabulário, as palavras devem ser registadas juntas (tradução nossa).

[27]Lewis (2000) aponta como estratégia para resolver os problemas de desmotivação dos alunos na aprendizagem da língua a realização de atividades promotoras da aprendizagem de *chunks - extensões* da língua*,* os quais integram *collocations*, *phrasal verbs* e *idioms*.

a língua por meio das suas próprias observações, transformando-se em agente do seu processo de aprendizagem.

Segundo Sardinha (2004, p. 292), as principais vantagens da abordagem *DDL – Data Driven Learning* giram em torno do facto de os alunos desenvolverem a aptidão de identificar regularidades e de realizar generalizações para as explicar. Por outro lado, o professor adota o lugar de guia ou mediador da pesquisa. Além disso, o ensino da gramática deixa de ser baseado na transmissão de regras.

A lexicometria é um procedimento metodológico e tecnológico de natureza objetiva, descritiva, indutiva e científica, que trata estatisticamente dados qualitativos sob fundo quantitativo, no sentido de caracterizar o contexto e a combinação de elementos lexicais de um determinado corpus. A lexicometria apresenta-se como um

> conjunto de métodos que permitem operar, a partir de análises estatísticas, reorganizações formais do vocabulário (conjunto de formas atualizadas no discurso, atestadas num texto ou num corpus de textos). O estudo lexicométrico impõe o levantamento exaustivo de TODAS as ocorrências, de TODAS as formas do corpus a estudar (Carvalho et al., 1999, p. 225).

A análise estatístico-lexical, além de nos permitir aceder a um inventário rigoroso e minucioso do vocabulário de determinado corpus, fornece-nos também, devido aos programas de análise estatística, resultados sistematizados e objetivos, assegura-nos a distância necessária entre o corpus e o investigador, contribuindo, desse modo, para uma exposição objetiva e neutra dos dados quantificados. Neste sentido, concordamos com Santana (1995, p. 4), ao considerar que, com este tipo de metodologia, "os alunos das tradicionais 'Letras' podem executar trabalhos tão válidos e científicos como os seus colegas das, também, tradicionais 'ciências exatas'. Penso que não será de menosprezar uma tal proposta".

Equacionada sob um certo ponto de vista de configuração tradicionalista, a utilização da informática na análise lexical afigura-se improfícua, no entanto, muitos estudiosos das Humanidades em geral, para além de revelarem a salutar consciência da indispensável adesão das Humanidades à informática, como forma de garantir a vitalidade das Humanidades, no que respeita à análise estatístico-lexical, preconizam que a utilização do computador constitui uma mais-valia:

> The computer is capable of processing vasts amounts of material in a very short time, and with total accuracy. It can do in a few moments some things that it would take a human being many days or even weeks to do. Furthermore, the computer can produce information from texts in a form which reveals patterns that a human reader would probably never even notice (Wray et al., 1998, p. 213)[28].

Através da informática, é-nos possível observar, de forma rigorosa, a frequência com que determinadas palavras ocorrem no texto, analisar as palavras-

[28]O computador é capaz de processar grandes quantidades de material num tempo muito curto e com total precisão. Pode fazer, em alguns momentos, algumas coisas que a um ser humano levaria dias ou mesmo semanas a fazer. Além disso, o computador pode produzir informações de textos de uma forma que revele padrões que um leitor humano provavelmente nem perceberia (tradução nossa).

chave, as palavras-tema, as formas exclusivas ou formas de frequência 1, os *hapax legomena* ou *hapaxes* (Hunston 2006), entre outros aspetos.

Biber (2011), referindo a integração dos métodos da linguística de corpus e dos objetivos e métodos de pesquisa da estilística tradicional, considera que, ao longo das últimas décadas, os métodos de pesquisa baseados em corpus têm sido usados para estudar a literatura e tornaram-se cada vez mais populares:

> In sum, it can be argued that corpus-based research methods have been used to study literature for the past 50 years or more. Over the past decade, this research has become increasingly popular, carried out under the umbrella of 'corpus stylistics'. These recent studies are explicit in their goals of integrating both the methods of corpus linguistics and the goals and methods of traditional stylistic research. However, there seems to be great potential for new lines of research that integrate the statistical methods of earlier research with the more rhetorical concerns of recent studies (2011, p. 22)[29].

4 CONCORDÂNCIAS E SUA UTILIZAÇÃO NA SALA DE AULA

A concordância fornece todos os contextos de ocorrência dos itens linguísticos do corpus. Por conseguinte, é possível verificar se determinado contexto em que certo item lexical ocorre é uma colocação, ou seja, se é frequente determinados itens surgirem contíguos em contextos linguísticos específicos.

A leitura das concordâncias deve ser efetuada a partir da coluna do nódulo, em direção aos seus colocados, tanto para a direita como para a esquerda. Sardinha (2004, p. 272) sublinha que, desse modo, é possível identificar "o ambiente mais típico do nódulo", acedendo aos seus diversos sentidos/significados e respetivos padrões de uso.

De facto, as definições de concordância encontram-se completamente ligadas ao conceito de colocações/colocados. Uma definição de colocação implica fazer referência a Firth, que foi o primeiro a usar o termo e a explicá-lo com a sua célebre frase: "you shall judge a word by the company it keeps" (1957b, p. 11)[30]. O autor considerou que os interlocutores, no ato de interação comunicativa real, possuem expetativas relativamente ao modo de utilização dos itens linguísticos, isto é, preveem os cotextos associados a determinados itens linguísticos. Por isso, conclui-se que a linguagem se baseia em padrões léxico-gramaticais, sabidos e partilhados pelos falantes de um mesmo idioma. De acordo com Firth, "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea

[29]Em suma, pode-se argumentar que os métodos de pesquisa baseados em corpus foram utilizados para estudar literatura nos últimos 50 anos ou mais. Na última década, essa pesquisa tornou-se cada vez mais popular, realizada sob a égide da '*corpus stylistics*'. Esses estudos recentes são explícitos nos seus objetivos de integrar os métodos da linguística de corpus e os objetivos e os métodos da pesquisa estilística tradicional. No entanto, parece haver um grande potencial para novas linhas de pesquisa que integram os métodos estatísticos de pesquisas anteriores com as preocupações mais retóricas de estudos recentes (tradução nossa).

[30]Deves julgar uma palavra pela companhia que mantém (tradução nossa).

approach to the meaning of words. One of the meanings of night is its collocability with dark, of course, collocation with night" (1957a, p. 196)[31].

Partington (1998, p. 16-17), partindo da definição de colocação de Sinclair (1991), de Leech (1974) e de Hoey (1991), apresenta três distintas definições de colocação:

> (1) Textual: "Collocation is the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991, p. 170)[32].
>
> (2) Psicológica: "Collocative meaning consists of the associations a word acquires on account of the meanings of words which tend to occur in its environment" (Leech, 1974, p. 20)[33].
>
> (3) Estatística: "collocation has long been the name given to the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, p. 6-7)[34].

A utilização de concordâncias na aula de língua é extremamente profícua. No âmbito dessa proficuidade, Sardinha (2004, p. 279) salienta "a obtenção de respostas a perguntas" não respondidas em obras didáticas, o "desenvolvimento do espírito pesquisador", "a independência em relação ao professor", o "incentivo à postura ativa do aluno" e o "centramento no aluno e a individualização do aprendizado". No entanto, o autor refere alguns impedimentos à utilização de concordâncias, sobretudo o facto de esse recurso didático exigir que os alunos e os professores sejam consciencializados e exercitados para a leitura direcionada pela palavra de busca (nódulo), tendo em vista a análise dos seus significados/sentidos e usos.

Através da análise de concordâncias na sala de aula, os alunos têm a oportunidade de confirmar as regras da gramática ou o uso e as especificidades léxicas, analisando palavras-chave em determinado contexto. Estas atividades implicam os alunos no processo didático, levando-os a desempenhar um papel mais ativo na aprendizagem do seu próprio vocabulário e, em função do seu nível de ensino, podem até discutir algumas das regras, a partir das suas observações dos padrões da língua. Entre outras atividades, os alunos podem descobrir formas linguísticas e significados novos, observar colocações típicas, relacionar palavras sintaticamente, identificar casos linguísticos de ambiguidade e polissemia, entre outros.

Efetivamente, as vantagens da utilização de *corpora* no processo de ensino/aprendizagem de uma língua residem na possibilidade que o principiante tem de aceder a um manancial muito amplo de exemplos de uso real da língua.

---

[31]O significado de colocação é uma abstração no nível sintagmático e não está diretamente preocupado com a abordagem conceptual ou a ideia do significado das palavras. Um dos significados da noite é a sua colocação com o escuro, é claro, colocação com a noite (tradução nossa).

[32]Colocação é a ocorrência de duas ou mais palavras num espaço curto num texto (tradução nossa).

[33]A colocação consiste nas associações que uma palavra adquire, devido aos significados das palavras que tendem a ocorrer em seu ambiente (tradução nossa).

[34]Colocação, há muito tempo, é o nome dado ao relacionamento que um item lexical possui com itens que aparecem com maior probabilidade no seu contexto (textual) (tradução nossa).

5 CONCLUSÕES

A Linguística de Corpus revolucionou o pensamento e atuação linguísticos da atualidade. Hoje, com as ferramentas, cada vez mais completas, diversificadas e de fácil acesso, que a Linguística Computacional e de Corpus coloca ao dispor de quantos se interessam pelo estudo da língua, as possibilidades em aberto são a cada dia maiores e mais abrangentes, ao nível da pesquisa, investigação e ensino. Com efeito, já ninguém nega que o contributo da informática para a elaboração, a edição e o estudo de textos é hoje indiscutível, e que esse contributo tem acompanhado o desenvolvimento técnico do computador e do *software*.

A dificuldade que envolve a definição de Linguística de Corpus, como verificámos, como uma teoria ou como uma metodologia, tem sido debatida a partir de diferentes pontos de vista. Tem sido argumentado que a Linguística de Corpus não é verdadeiramente um domínio de pesquisa, mas apenas uma base metodológica para estudar a linguagem. No entanto, muitos linguistas que trabalham com corpus tendem a concordar que a Linguística de Corpus vai muito para além desse papel exclusivamente metodológico, visão que partilhamos.

O trabalho com *corpora* na sala de aula acarreta uma aproximação entre as práticas de investigação e as práticas de ensino-aprendizagem. O aluno adquire o papel de um investigador que pretende obter respostas a partir dos dados disponíveis no corpus. Deste modo, o aluno descobre a língua por meio das suas próprias observações, transformando-se em agente do seu processo de aprendizagem.

Através da informática, é-nos possível observar, de forma rigorosa, a frequência com que determinadas palavras ocorrem no texto, analisar as palavras-tema, as formas exclusivas ou formas de frequência, utilizar concordâncias, entre outros aspetos.

Estas atividades implicam os alunos no processo didático, levando-os a desempenhar um papel mais ativo na aprendizagem do seu próprio vocabulário e, em função do seu nível de ensino, podem até discutir algumas das regras, a partir das suas observações dos padrões da língua. Entre outras atividades, os alunos podem descobrir formas linguísticas e significados novos, observar colocações típicas, relacionar palavras sintaticamente, identificar casos linguísticos de ambiguidade e polissemia, etc.. De facto, a Linguística de Corpus proporciona ao ensino progressos inegáveis. Por isso, a este trabalho seguir-se-á um estudo, utilizando o programa NooJ, sobre os contos da tradição oral portuguesa, dando enfoque especial às concordâncias e à frequência das palavras e seu contributo para o ensino.

REFERÊNCIAS

Biber D. Corpus linguistics and the study of literature: back to the future? Scientific Study of Literature, 2011;1(1):15–23.

Bowkler L, Jennifer P. Working with specialized language: a practical guide to using corpora. London/New York: Routledge; 2002.

Carvalho D, Marques MER, Silva MF. Discurso: práticas lexicométricas. Linguística computacional: investigação fundamental e aplicações. Lisboa: Edições Colibri/Associação Portuguesa de Linguística; 1999.

Chomsky N. Three models for the description of language. IRE transactions on information theory IT-2; 1956;2(3)sept.:113-124.

Chomsky N. The logical basis of linguistic theory. In: Proceedings of the 9th international congress of linguists. Horace Lunt; 1962. p. 914-978. The Hague: Mouton; 1964.

Firth JR. Modes of meaning. In: Firth JR, editor. Papers in linguistics 1934-1951. London: Oxford University Press; 1957a. p. 190-215.

Firth JR. A synopsis of linguistic theory 1930–1955. In: Studies in linguistic analysis. Philological Society. 1957b;special volume:1–32.

Gries S. Corpus linguistics and theoretical linguistics: A love–hate relationship? Not necessarily. In: International Journal of Corpus Linguistics. 2010;2(3):327–343. [citado 11 nov. 2019]. Disponível em: http://www.linguistics.ucsb.edu/faculty/stgries/research/2010_STG_CorpLingLingTheory_IJCL.pdf.

Halliday MAK. Corpus studies and probabilistic grammar. In: Aijmer K, Altenberg B, organizadores. English corpus linguistics: studies in honour of Jan Svartvik. London: Longman; 1991. p. 30-43.

Halliday MAK. Computational and quantitative studies. Edited by Jonathan J. Webster. London: Continuum; 2006.

Hoey M. Patterns of lexis in text. Oxford: Oxford University Press; 1991.

Hunston S. Corpora in applied linguistics. Cambridge: Cambridge University Press; 2006.

Johns T. From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In: Jonhs T, King P, editores. Classroom concordancing. ELR Journal. Birmingham University Press. 1991;4:27-46.

Kennedy G. An introduction to corpus linguistics. London: Longman; 1998.

Leech, G. Semantics. Harmondsworth: Penguin Books; 1974.

Leech G. The state of the art in corpus linguistics. In: Aijmer K, Altenberg B, editores. English Corpus Linguistics: Studies in honour of Jan Svartvik. London: Longman; 1991. p. 20-41.

Leech G. Corpora and theories of linguistic performance. In: Svartvik J, editor. Directions in corpus linguistics. Proceedings of nobel symposium, 4-8 august 1991. Berlin/New York: Mouton de Gruyter; 1992. p. 105-122.

Leech G. Teaching and language corpora: a convergence. In: Wichmann A, et al., editores. Teaching and language corpora. New York: Longman; 1997. p 1-24.

Lewis M. Language in the lexical approach. In: Michael L, editor. Teaching collocation: further developments. The lexical approach. Hove: Language Teaching Publications; 2000. p. 126-154.

McEnery T, Wilson A. Corpus linguistics. Edinburgh: Edinburgh University Press; 1996.

McEnery T, Gabrielatos C. English corpus linguistics. In: Aarts B, Mcmahon A, editores. The handbook of english linguistics. Oxford: Blackwell; 2006. p. 33-71.

McEnery T, Xiao R, Tono Y. Corpus-based language studies: an advanced resource book. London/New York: Routledge; 2006.

Mello H, Sousa R. A linguagem da ciência: prospeção de dados baseados em corpora. In: Anais – Seminários teóricos interdisciplinares do SEMIOTEC – I STIS; 2012. p. 1-19.

Meyer CF. English corpus linguistics: an introduction. Cambridge: Cambridge University Press; 2002.

Partington A. Patterns and meanings – using corpora for English language research and teaching (studies in corpus linguistics 2). Amsterdam/Philadelphia: John Benjamins; 1998.

Sardinha TB. Lingüística de corpus: histórico e problemática. Delta. 2000;16(2):323-367.

Sardinha TB. Lingüística de corpus. São Paulo: Manole; 2004.

Sinclair J. Corpus, concordance, collocation. Oxford: Oxford University Press; 1991.

Sinclair J. Trust the text language, corpus and discourse. Edited with Ronald Carter. London: Routledge; 2004.

Stubbs M. British traditions in text analysis: from Firth to Sinclair. In: Baker M, Francis F, Tognini-Bonelli E, editores. Text and technology: in honour of John Sinclair. Amsterdam/Philadelphia: John Benjamins; 1993. p. 1–46.

Teubert W. My version of corpus linguistics. In: International Journal of Corpus Linguistics. 2005;10 (1):1–13. [citado 11 nov. 2019]. Disponível em: http://www.corpus4u.org/forum/upload/forum/2005071006505939.pdf.

Teubert W. Rethinking corpus linguistics. In: A mosaic of corpus linguistics: selected approaches. Sánchez A, Almela M, editores. Frankfurt: Internationaler Verlag der Wissenschaften; 2010. p. 19-41.

Tognini-Bonelli E. Corpus linguistics at work. Amsterdam/Philadelphia: John Benjamins; 2001.

Willis D. The lexical syllabus: a new approach to language teaching. London: Collins ELT; 1990.

Wray A, Trott K, Bloomer A. Projects in linguistics – a practical guide to researching language. London: Arnold; 1998.