DOI: 10.14393/DL52-v16n4a2022-4

O fenômeno do desfocamento do agente: uma discussão sobre a importância dos recursos computacionais para os estudos linguísticos

The phenomenon of agent defocusing: a discussion of the relevance of computational resources for linguistic studies

Andressa Rodrigues GOMIDE[*]
Taíse SIMIONI[**]
Aden PEREIRA[***]

RESUMO: Embora em expansão, a pesquisa linguística empírica da língua portuguesa ainda está longe de alcançar todo o seu potencial. Acreditamos que isso possa se dever, em parte, pelo desconhecimento de alguns investigadores de recursos já disponíveis gratuitamente. Neste artigo, apresentamos algumas ferramentas da Linguística de *Corpus* e um *corpus* de escrita acadêmica em português (CoPEP), e como eles podem ser utilizados para explorar o fenômeno do desfocamento do agente em artigos acadêmicos publicados no Brasil e em Portugal. Para isso, utilizamos recursos já existentes para anotar e disponibilizar de forma gratuita e online o CoPEP, um *corpus* de extrema utilidade para investigações linguísticas acerca do português acadêmico.

PALAVRAS-CHAVE: Linguística de *Corpus*. Escrita acadêmica. Ferramentas de investigação de *corpora*. Desfocamento do agente.

ABSTRACT: Although expanding, empirical linguistic research on the Portuguese language is still far from reaching its full potential. We believe that this might be due to some researchers lack of awareness of resources already available for free. In this article, we present some Corpus Linguistics tools and a corpus of academic writing in Portuguese (CoPEP), and how they can be used to explore the phenomenon of agent defocusing in academic articles published in Brazil and Portugal. For this, we use existing resources to annotate and make available, free of charge and online, the CoPEP, an extremely useful corpus for linguistic investigations on academic Portuguese.

KEYWORDS: Corpus Linguistics. Academic writing. Tools for corpus analysis. Agent defocusing.

[*] Doutora. Universidade de Coimbra. ORCID: https://orcid.org/0000-0002-1481-4748. andressa.gomide@fl.uc.pt
[**] Doutora. Universidade Federal do Pampa. ORCID: https://orcid.org/0000-0002-9778-7393. taisesimioni@unipampa.edu.br
[***] Doutora. Universidade Federal do Pampa. ORCID: https://orcid.org/0000-0002-2866-4218. adenpereira@unipampa.edu.br.

1 Introdução

A quantidade de recursos computacionais (dados e ferramentas) disponíveis ou em fase de desenvolvimento no âmbito do processamento computacional da língua portuguesa está em franco crescimento. Tal fenômeno é observado tanto na indústria quanto na academia. Na indústria, possivelmente isso é causado pelo aumento da classe consumidora de tecnologias em Estados membros da Comunidade dos Países de Língua Portuguesa (CPLP). Há um crescente interesse em desenvolvimento de produtos como assistentes de voz (ex.: Alexa, Cortana, Siri) e *chatbots*. Na academia, temos grandes projetos em andamento, como a criação do primeiro Dicionário do Português de Moçambique (DiPoMo)[1] e o projeto de Reconhecimento Automático de Fala e Síntese de Fala no Centro de IA (TaRSila)[2].

Contudo, o progresso no contexto acadêmico não é tão rápido como gostaríamos. Há ainda um número reduzido de estudos linguísticos empíricos que utilizam métodos e dados computacionais. Acreditamos que uma razão para o número reduzido de estudos linguísticos computacionais do português não está na escassez de ferramentas e dados, mas no desconhecimento de tais ferramentas, ou mesmo, na insegurança por parte dos pesquisadores em usá-las.

Com esta suposição em mente, temos como objetivos deste artigo (a) traçar um breve histórico da Linguística de *Corpus* e discutir suas contribuições para análises linguísticas (seção 2), (b) apresentar dois estudos empíricos feitos a respeito do fenômeno do desfocamento do agente no português brasileiro (PB) (seção 3), (c) e demonstrar como podemos expandir as pesquisas apresentadas em (b) com o uso de um *corpus* de escrita acadêmica junto ao uso de uma ferramenta de busca em *corpus* e à aplicação de testes de análise estatística.

---

[1] Mais informações disponíveis no endereço eletrônico: https://www.instituto-camoes.pt/sobre/comunicacao/noticias/mocambique-projeto-do-primeiro-dicionario-de-portugues-de-mocambique-arranca-com-formacao-em-maputo

[2] Mais informações disponíveis no endereço eletrônico: https://sites.google.com/view/tarsila-c4ai

2 Relevância da Linguística de *Corpus* para análises quali-quantitativas

Segundo Berber Sardinha (2004, p. xvii), a Linguística de *Corpus* (LC) é uma área que trata do uso de *corpora* computadorizados que se compõem de textos coletados, transcrições ou escritos da fala, sendo estes mantidos em arquivo de computador. Assim, a LC busca contestar os paradigmas linguísticos em prol de abrir novos caminhos para diversos estudiosos da área da linguagem, tais como linguistas, professores, tradutores e lexicógrafos, a título de exemplo, bem como outros profissionais, beneficiando essas áreas de modo a instrumentalizá-las quanto às pesquisas das línguas que o investigador ou investigadora pode realizar em alinhamento com a LC.

De acordo com Berber Sardinha (2004, p. 4), o primeiro *corpus* linguístico eletrônico, o *Brown University Standard Corpus of Present-day American English*, criado em 1964, contava com 1 milhão de palavras. A partir dele, os textos eram transferidos para o computador por meio de cartões perfurados um a um. O autor ainda acrescenta que o primeiro *corpus* de língua falada era composto por 200 mil palavras, as quais foram coletadas pelo estudioso Sinclair (1995).

Naquele período, a coleta de dados linguísticos era vista com desconfiança na academia ao mesmo tempo em que Chomsky lançava a Teoria do Gerativismo, com maior destaque ao que a mente podia processar para, a seguir, ser convertido em linguagem, ou seja, a ênfase era dada mais à competência do que ao desempenho do falante, segundo Berber Sardinha (2004).

No Brasil, em meados dos anos 2000, a LC estava em fase preliminar, voltada mais para a Lexicografia e a Linguística Computacional, ainda que entre essas áreas não houvesse conformidade recíproca sobre qual seriam as funções da LC, seja para coleta, análise e processamento dos dados, os quais poderiam beneficiar tais áreas como instrumentos de pesquisa, como esclarece Berber Sardinha (2004, p. 6).

Ainda assim, com o crescimento da busca por mais produtos tecnológicos que propiciem a melhoria da comunicação, tanto quanto a de bens e produtos, ademais da esfera acadêmica, foi perceptível o crescimento do interesse no meio corporativo sobre *corpus*. Desse modo, muitas parcerias entre as universidades e as empresas sucederam-se.

Em tempos mais recentes, vimos testemunhando o interesse progressivo no processamento automático de textos, assim como na informatização de bases de dados e na montagem de sistemas inteligentes de reconhecimento de voz e gerenciamento da informação. Isso, especialmente, em se tratando de empresas de telecomunicações e marketing digital que buscam investir nesses elementos característicos da área de computação.

Uma das ferramentas bastante utilizadas para o processamento de *corpora* linguísticos é o programa *WordSmith Tools*, que apresenta diversas versões (gratuitas e pagas). A partir desse recurso, pode-se investigar a frequência, os graus de fixidez e decomponibilidade que ocorrem nas palavras, expressões e sentenças dos mais variados gêneros textuais.

Berber Sardinha (2004, p. 22) destaca a importância da representatividade de um *corpus* em uma língua. Nesse mesmo sentido, Leech (1992, p. 120) traz o foco para as investigações de Biber, de modo a apontar a função representativa de um *corpus*, já que aquela tem o papel de caracterizá-lo, haja vista que "No design de corpus, a representatividade da linguagem é alcançada por progressão cíclica, empiricamente a partir do teste de adequação de corpora previamente elaborados" (p. 120)[3].

Assim, Leech (1999) e Sinclair (1995) entendem que, para se ter um *corpus* representativo, é fundamental o conhecimento acerca da população de onde ele é derivado, porque isso estaria correlacionado à possibilidade de se estabelecer

[3] Texto original: "[...] in corpus design, representativeness of the language is achieved by cyclic progression, based on empirically testing the adequacy of previously designed corpora".

associações entre traços que parecem ser mais ou menos comuns em certos cenários. Em outros termos, para Berber Sardinha (2004, p. 22), conhecer a probabilidade em que ocorrem os traços lexicais, estruturais, pragmáticos e discursivos nos mais diversos contextos está no âmago da LC.

Além da representatividade apontada por Berber Sardinha, segundo Rocha (2007, p. 20), há outras características que um *corpus* precisa apresentar a fim de ser constituído como tal: amostra, tamanho finito e formato legível em um computador e referência padrão. Isso porque, para esse autor, "[...] a representatividade determina quais generalizações em relação às características de uma determinada população são confiáveis, muitas vezes expressa em termos de populações às quais as generalizações se aplicam" (p. 21) [4].

Rocha (2007) destaca que Biber *et al.* (1998) debatem a noção de *corpus* equilibrado, categorizando aquilo que deve ser incorporado ao próprio *corpus* porque para o autor "[...] as técnicas típicas de amostragem utilizadas em estudos estatísticos só são úteis à linguística de forma limitada" [5], já que, por exemplo, "[...] uma amostra proporcional de uma língua, tal como registrada através de um grupo de usuários da língua em suas atividades diárias, resultaria em um *corpus* homogêneo" [6].

Por outro lado, excluir tais textos poderia comprometer a representatividade, já que eles são gêneros utilizados com frequência na sociedade (ROCHA, 2007). Assim, no presente artigo, buscou-se trabalhar com um *corpus* representativo do gênero acadêmico para que se pudesse mensurar com maior precisão a ocorrência do fenômeno do desfocamento do agente.

---

[4] Texto original: "[…] representativeness determines which generalisations regarding features of a given population are trustworthy, often expressed in terms of populations to which generalisations apply".
[5] Texto original: "[…] typical sampling techniques used in statistical studies are only useful to linguistics to a limited extent".
[6] Texto original: "[…] a proportional sample of a language, as registered through a group of language users in their daily activities, would result in a rather homogeneous corpus"

Atualmente, os estudos de Biber destacados por Berber Sardinha e Pinto (2014), a partir do 25º aniversário da publicação do livro seminal daquele autor, *Variation Across Speech and Writing*, em 1988, apontam para o estudo de Análise Multifuncional (AM), já que diversos pesquisadores têm como proposta ampliar o escopo da AM analisando um espectro de registros, períodos de tempo e contextos de uso dos diversos *corpora* existentes desde aquele período histórico do surgimento da LC.

A flexibilidade da abordagem da AM torna possível destacar a eficácia de sondar tanto contextos especializados quanto os mais gerais com eficiência expressiva, além de aferir o que já existe há anos ou décadas de uso da linguagem. Neste sentido, de acordo com Halliday (1993/2005 *apud* BERBER SARDINHA; PINTO, 2014), já é possível detectar uma gama de probabilidades de uso de uma língua ou de suas variedades em diversos contextos, levantando hipóteses contundentes resultantes das mais diversas experiências humanas e de suas interações comunicacionais a partir dos experimentos da AM.

Para Berber Sardinha e Pinto (2014), a AM é um método muito potente que permite ao pesquisador empreender uma análise da língua em uso, a partir da qual é possível fazer descrições que consigam capturar como os usuários da língua fazem suas escolhas linguísticas em contextos específicos, já que a AM se fundamente na vida real humana. No caso do presente artigo, é possível verificar as escolhas dos acadêmicos ao se referirem aos seus objetos de estudo e verificar como fazem essas escolhas a partir dos contextos que o *corpus* nos apresenta.

Assim, em trabalhos futuros, será possível investigar, na sequência da discussão feita no presente artigo, conforme apontam Berber Sardinha e Pinto (2014),

> relações intensas com a comunicação humana, observação atenta do contexto e das pessoas que vivem nas situações em que a linguagem é usada e o movimento analítico rápido de um registro para outro em

> um esforço para perceber as qualidades que os unem ou os separam (p. xv).[7]

Isso sempre levando-se em consideração que o olhar investigativo de quem analisa o objeto de estudo, por mais objetivo que seja o método, sempre parte de sua perspectiva vivencial também como usuário da língua, no caso, o português do Brasil aqui em questão.

Após essa breve discussão sobre pressupostos importantes para a LC, a próxima seção apresenta os resultados de dois estudos sobre o fenômeno do desfocamento do agente, o que nos permitirá, na seção 4, apontar para caminhos possíveis de expansão de tais pesquisas a partir da exploração de recursos computacionais.

3 O fenômeno do desfocamento do agente

Morais (2016) analisa o uso do clítico *se* em artigos acadêmicos a fim de verificar a sua atuação como uma estratégia de impessoalização ou de desfocamento do agente. O *corpus* de sua pesquisa faz parte do projeto SAL (*Systemics Across Languages*) e foi constituído por 1225 artigos produzidos em língua portuguesa e coletados na plataforma *Scielo*. A autora destaca o uso de ferramentas computacionais em seu trabalho, especificamente o programa *WordSmith Tools*, o que lhe permitiu a análise de uma grande quantidade de textos produzidos em situações reais de interação.

Como etapa metodológica, Morais (2016) organiza os diferentes usos do clítico *se* em grupos. Para tal, a autora se utiliza de testes de refraseamento. Como o fenômeno que nos interessa no presente trabalho é o do desfocamento do agente, nossa atenção se voltará para o que a autora classificou como grupo 2. Trata-se de usos do clítico *se*

[7] Texto original: "[...] intense dealings with human communication, thoughtful consideration of the context and the people living in those situations where the language is used, and swift analytical movement from one register to the next in an effort to perceive the qualities that bond them together or tease them apart"

que equivalem a construções com passiva analítica ou com primeira pessoa do plural. Os exemplos em (1) mostram tais equivalências[8].

(1)
(a) Observa-se que houve diferença significativa…
(b) Foi observado que houve diferença significativa…
(c) Observamos que houve diferença significativa…

Como esclarece Morais (2016), neste grupo, há uma predominância de processos materiais, na terminologia usada pela linguística sistêmico-funcional, perspectiva teórica adotada pela autora. As equivalências em (1) evidenciam a existência de um agente, de forma que, na construção (1a), "o clítico é um mecanismo importante para *apagar* o autor no texto, deixando, porém, um resquício de sua participação" (MORAIS, 2016, p. 79).

Este grupo é incluído por Morais (2016) na categoria em que o clítico *se* encontra-se em construções com desfocamento de participante. Morais (2013) propõe três graus de desfocamento de participantes: no alto grau, qualquer participante poderia estar envolvido; no médio grau, dois participantes estariam envolvidos (o autor do artigo e a comunidade acadêmica); no baixo grau, haveria um único participante envolvido (o autor do artigo ou o pesquisador que estiver sendo mencionado).

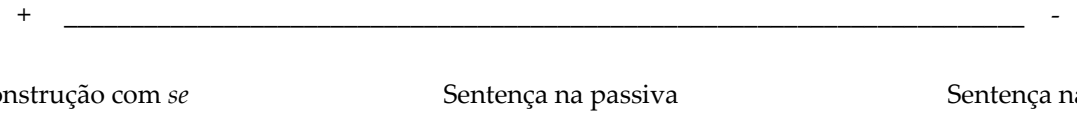
O presente estudo tem como foco o que Morais (2013) classifica como baixo grau de desfocamento do participante, construções nas quais é comum a presença de circunstâncias de lugar e a ocorrência de verbos no pretérito, como exemplifica a sentença em (2).

(2)
Nessa pesquisa verificou-se valor de produção de matéria seca…

[8] Os exemplos em (1) e (2) foram retirados de Morais (2016).

Como explica Morais (2016, p. 91), "o artigo de pesquisa é um texto em que se relata sobre uma pesquisa feita, por isso, os processos ligados ao processo de *fazer* pesquisa (*observar*, *verificar* e *analisar*) permitem pressupor um Agente, o pesquisador, responsável pelas etapas/ações do trabalho". Partindo desse pressuposto, propomos que as sentenças em (1) podem revelar um contínuo de desfocamento do agente, como mostra Fig. 1.

Figura 1 – Contínuo de desfocamento do agente.



Fonte: elaborado pelas autoras.

No polo à esquerda, estaria localizada a construção com *se*, que permite o grau máximo de desfocamento, como aquela em (1a). No polo à direita, estariam as construções ativas, em que não ocorre o desfocamento do agente, como aquela em (1c). Entre os polos, se encontrariam as construções na voz passiva, que permitiriam um grau intermediário de desfocamento, como aquela em (1b).

Tal contínuo baseia-se na análise de Shibatani (1985), segundo a qual a função primária das construções passivas é o desfocamento do agente (*agent defocusing*). Segundo o autor, tal função se evidencia pelo fato de que as construções passivas geralmente não expressam seus agentes de forma explícita, mesmo nas línguas que permitem tal estrutura. Camacho (2002), que analisou um *corpus* constituído por 916 ocorrências de estruturas sentenciais pertencentes ao NURC (Projeto da Norma Urbana Culta), mostra que este é o caso do português brasileiro. Segundo seus resultados, em 85,5% das ocorrências de passivas, não existe "a possibilidade de

recuperação, no contexto discursivo, de referência a uma entidade individuada que seja controladora da ação desenvolvida no predicado" (CAMACHO, 2002, p. 258).

Shibatani (1985), ao defender a sua proposta de análise das passivas como estruturas que têm por finalidade o desfocamento do agente, propõe a seguinte hierarquia para o foco no que diz respeito à estrutura de uma sentença: sujeito > objeto direto > objeto indireto > objetos oblíquos. Em línguas como o português, que permitem a expressão do agente em construções passivas, o agente é expresso por um objeto oblíquo, ou seja, ele assume "o grau mais baixo de foco entre os elementos sintaticamente codificados"[9] (SHIBATANI, 1985, p. 833).

No contínuo de desfocamento do agente proposto neste trabalho, situamos as construções com *se* no grau máximo de desfocamento, distinguindo-as das passivas, em função de que, como mostra Camacho (2002, p. 251), as primeiras "não autorizam a manifestação formal de um SN agentivo", como mostram os exemplos em (3)[10].

(3)
(a) João quebrou o vidro da janela.
(b) O vidro da janela foi quebrado (por João).
(c) O vidro da janela (se) quebrou (?por João).

Camacho (2002) explica que a diferença mostrada entre (3b) e (3c), que reside na possibilidade ou não de expressão de um agente, é tão relevante que isso se relaciona à possibilidade de haver uma expressão de instrumento nas passivas (4a), o que pressupõe a existência de uma entidade agentiva, enquanto as construções com *se* não a permitiriam (4b).

(4)
(a) O vidro da janela foi quebrado com uma pedrada.

[9] Texto original: "[...] the lowest degree of focus among the syntactically encoded elements".
[10] Os exemplos em (3) e (4) foram retirados de Camacho (2002).

(b) O vidro da janela (se) quebrou (?com uma pedrada).

Entre as características comuns à construção passiva e àquela com *se*, Camacho (2002) cita o fato de que ambas acarretam um argumento agentivo, ainda que na segunda tal argumento fique subentendido, uma vez que inexiste a possibilidade de sua manifestação formal. Ainda conforme o autor, "mesmo que, na passiva, o agente nem sempre se manifeste, enunciá-lo depende unicamente do ponto de vista do falante em relação ao evento e não de uma restrição sintático-semântica" (CAMACHO, 2002, p. 304), diferentemente do que ocorre nas construções com *se*, nas quais o agente não pode ser expresso.

É com base nestes fatores que propomos a diferença entre as construções com *se* e as passivas no que diz respeito ao seu lugar no contínuo de desfocamento do agente.

Destacamos que o presente trabalho é derivado de uma reflexão sobre a língua que foi possível a partir da análise dos resultados de uma pesquisa que se utilizou de ferramentas computacionais, o que permitiu o acesso a uma grande quantidade de textos (MORAIS, 2016), combinada com a análise de resultados de uma pesquisa que, apesar de não se valer de ferramentas computacionais, procedeu a um escrutínio criterioso dos dados (CAMACHO, 2002). Tal percurso teórico-metodológico nos permitiu apontar possibilidades de expansão do estudo do desfocamento do agente a partir da exploração de recursos computacionais, conforme mostra a seção 4 deste trabalho.

4 O desfocamento do agente em um *corpus* de escrita acadêmica

A seção anterior apresentou estudos que discutem as diferentes formas de desfocamento do agente e suas implicações. Embora os estudos apontem interessantes resultados, a pesquisa empírica com um alto volume de dados somada ao uso de

ferramentas computacionais adequadas pode enriquecer os resultados. É tal ideia que será exposta e defendida nas próximas subseções.

4.1 O *corpus*

Para este estudo, utilizamos o *Corpus* de Português Escrito em Periódicos (CoPEP) (KUHN; FERREIRA, 2020). O CoPEP é um *corpus* representativo da escrita acadêmica em língua portuguesa nas variedades brasileira e europeia (PE), composto de 9.900 textos e um total de 48.506.519 palavras. O *corpus* reúne textos publicados entre os anos 1992 e 2018 em revistas acadêmicas periódicas de seis áreas de conhecimento, agrupadas em três grandes colégios (Tabelas 1 e 2).

Além do texto *per se*, podemos também obter informações adicionais para cada unidade mínima, *token*, de um *corpus*. *Tokens* são, em sua maioria, palavras. Dígitos, siglas, acrônimos, pontuação são também considerados *tokens*. Essas informações adicionais, também conhecidas como *anotação*, são normalmente feitas de forma automática. Atualmente uma versão do CoPEP com anotações para classes gramaticais (POS - do inglês *parts-of-speech*) e lema (ou *tokens* em sua forma neutra) está disponível via SketchEngine (KILGARRIFF *et al*., 2014). O SketchEngine é uma poderosa plataforma *online* de busca em *corpus*, disponibilizada por meio de assinatura paga.

Para o presente trabalho, obtivemos dos autores o *corpus* bruto, i.e., sem anotações linguísticas, e permissão para distribuí-lo gratuitamente. Do *corpus* em seu estado bruto, preparamos e anotamos um *corpus* com dois sistemas diversos para permitir que estudos futuros trabalhem com eles e testem diferentes tipos de anotações e suas respectivas precisões. O primeiro sistema possui anotações para classes gramaticais e lema, implementadas a partir do TreeTagger (SCHMID, 1994). O segundo sistema foi implementado utilizando o Spacy[11], uma biblioteca para

[11] Disponível em: https://spacy.io

processamento de linguagem natural em Python. Neste segundo sistema, temos as seguintes anotações: lema, classe gramatical, classe gramatical simplificada, núcleo do sintagma e etiqueta sintática.

Além das informações atribuídas a cada *token* do *corpus*, é útil também obter informações a respeito de cada texto que compõe o *corpus*. No CoPEP, as informações preservadas em cada texto são as seguintes: variedade do português (europeu ou brasileiro); país sede da revista acadêmica; área de conhecimento; grande área de conhecimento; ISSN; ano de publicação; título do artigo. Tais informações são comumente conhecidas como *metadados*. Para melhor explorar o potencial das anotações e metadados, precisamos instalar ou carregar o *corpus* em uma ferramenta de busca. Para o presente trabalho, instalamos o CoPEP em um aplicativo *online* de busca em *corpus*, o CQPweb (HARDIE, 2012)[12].

Tabela 1 – Distribuição de *tokens* e textos no CoPEP (área de conhecimento).

	Europa		Brasil		CoPEP Total	
	tokens	textos	tokens	textos	tokens	textos
Ciências Humanas	12747013	2581	12751623	1219	25498636	3800
Ciências Sociais Aplicadas	2686764	517	2689639	319	5376403	836
Colégio de Humanidades	15433777	3098	15441262	1538	30875039	4636
Ciências da Saúde	6687507	2432	6695452	1564	13382959	3996
Ciências Agrícolas	1283054	385	1301763	522	2584817	907
Colégio de Ciências da Vida	7970561	2817	7997215	2086	15967776	4903
Ciências Exatas e da Terra	402028	67	401918	118	803946	185
Engenharia	422790	107	436968	69	859758	176
Colégio de Ciências Exatas, da Terra e Multidisciplinar	824818	174	838886	187	1663704	361

Fonte: elaborada pelas autoras.

Tabela 2 – Distribuição de *tokens* e textos no CoPEP (ano de publicação).

	Todos		europeu		brasileiro	
	tokens	textos	tokens	textos	tokens	textos
1992	12496	1	0	0	12496	1
1993	29810	3	0	0	29810	3
1994	29198	3	0	0	29198	3
1996	21240	2	0	0	21240	2

[12] Disponível em: https://ola.unito.it/CQPweb32/copep

1997	89312	15	9343	9	79969	6
1998	367897	91	5048	6	362849	85
1999	270118	63	26332	17	243786	46
2000	654587	97	241728	31	412859	66
2001	797574	173	232095	37	565479	136
2002	1267869	278	203339	37	1064530	241
2003	1917247	407	63109	22	1854138	385
2004	1673253	368	65742	38	1607511	330
2005	1731566	330	121155	87	1610411	243
2006	2252933	443	300393	122	1952540	321
2007	1684884	330	371905	168	1312979	162
2008	1973120	363	664464	209	1308656	154
2009	3660192	684	2334265	530	1325927	154
2010	4246448	871	2508551	604	1737897	267
2011	4326782	910	2541318	647	1785464	263
2012	5534955	1146	3811356	903	1723599	243
2013	5710364	1191	3711249	910	1999115	281
2014	6037394	1298	3972501	1023	2064893	275
2015	3899954	769	2943845	664	956109	105
2016	281453	54	65545	15	215908	39
2017	28534	8	28534	8	0	0
2018	7339	2	7339	2	0	0
Total	48506519	9900	24229156	6089	24277363	3811

Fonte: elaborada pelas autoras.

4.2 Ferramenta de busca

Escolhemos o CQPweb por ser uma ferramenta *online* e de código aberto. Ser uma ferramenta *online* auxilia o estudo colaborativo e elimina a necessidade de instalação de programas. Possuir o código aberto significa, neste caso, ter uma ferramenta gratuita para o consumidor final e ainda ter a possibilidade de adaptar ou criar novas funções no programa, caso haja necessidade.

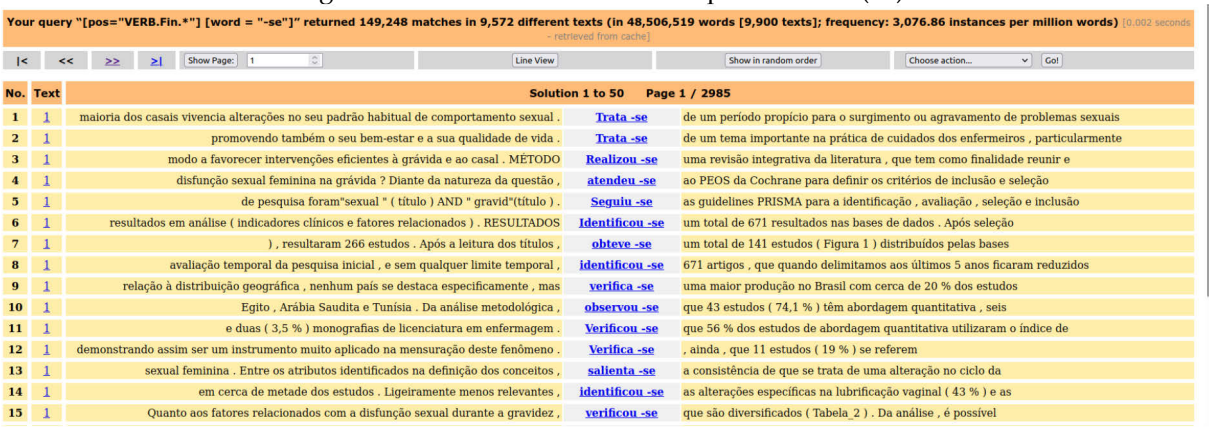
Com o CQPweb, é possível realizar buscas simples por palavras ou sequência de palavras e obter linhas de concordância, bem como a frequência relativa e absoluta dos elementos da busca e o número de textos que possuem o(s) item(ns) buscado(s). As buscas também podem ser restringidas (filtradas) de acordo com os metadados dos textos que compõem o *corpus*. Ou seja, é possível optar por fazer buscas apenas em textos com determinadas características.

Uma importante ferramenta do CQPweb é o poderoso sistema de busca *Corpus Query Processor* (CQP). Esse sistema permite que realizemos buscas refinadas por estruturas elaboradas. Por exemplo, para encontrarmos exemplos de desfocamento do agente, como descrição feita na seção três, podemos utilizar as formas de buscas em (5).

(5)
(a) [pos="VERB.Fin.*"] [word = "-se"]
(b) [(word = "foi") | (word = "foram")]{1,3} [(pos = "VERB.Part.*")]
(c) [(pos != "PRON.*") & (pos != "NOUN.*") & (word != "que")]{1,3} [(pos = "VERB.Fin.Plur") & (word = ".*mos")]

A busca (5a) procura por todas as ocorrências de um *token* com a etiqueta de verbo na forma finita seguido pelo *token* "-se" (Fig. 2). A estrutura em (5b) retorna ocorrências em que o *token* "foi" ou "foram" ocorre à esquerda de um verbo em sua forma participial (fig. 3). A notação {1,3} significa que o segundo *token* pode vir imediatamente após o primeiro (ex.: Foram analisados) ou até três casas à direita (ex.: Foi devidamente analisado).

Figura 2 – Linhas de concordância para a busca (5a).



Your query "[pos="VERB.Fin.*"] [word = "-se"]" returned 149,248 matches in 9,572 different texts (in 48,506,519 words [9,900 texts]; frequency: 3,076.86 instances per million words) [0.002 seconds - retrieved from cache]

|< << >> >| Show Page: 1 Line View Show in random order Choose action... Go!

No.	Text	Solution 1 to 50 Page 1 / 2985		
1	1	maioria dos casais vivencia alterações no seu padrão habitual de comportamento sexual .	Trata -se	de um período propício para o surgimento ou agravamento de problemas sexuais
2	1	promovendo também o seu bem-estar e a sua qualidade de vida .	Trata -se	de um tema importante na prática de cuidados dos enfermeiros , particularmente
3	1	modo a favorecer intervenções eficientes à grávida e ao casal . MÉTODO	Realizou -se	uma revisão integrativa da literatura , que tem como finalidade reunir e
4	1	disfunção sexual feminina na grávida ? Diante da natureza da questão ,	atendeu -se	ao PEOS da Cochrane para definir os critérios de inclusão e seleção
5	1	de pesquisa foram"sexual " (título) AND " gravid"(título) .	Seguiu -se	as guidelines PRISMA para a identificação , avaliação , seleção e inclusão
6	1	resultados em análise (indicadores clínicos e fatores relacionados) . RESULTADOS	Identificou -se	um total de 671 resultados nas bases de dados . Após seleção
7	1	) , resultaram 266 estudos . Após a leitura dos títulos ,	obteve -se	um total de 141 estudos (Figura 1) distribuídos pelas bases
8	1	avaliação temporal da pesquisa inicial , e sem qualquer limite temporal ,	identificou -se	671 artigos , que quando delimitamos aos últimos 5 anos ficaram reduzidos
9	1	relação à distribuição geográfica , nenhum país se destaca especificamente , mas	verifica -se	uma maior produção no Brasil com cerca de 20 % dos estudos
10	1	Egito , Arábia Saudita e Tunísia . Da análise metodológica ,	observou -se	que 43 estudos (74,1 %) têm abordagem quantitativa , seis
11	1	e duas (3,5 %) monografias de licenciatura em enfermagem .	Verificou -se	que 56 % dos estudos de abordagem quantitativa utilizaram o índice de
12	1	demonstrando assim ser um instrumento muito aplicado na mensuração deste fenômeno .	Verifica -se	, ainda , que 11 estudos (19 %) se referem
13	1	sexual feminina . Entre os atributos identificados na definição dos conceitos ,	salienta -se	a consistência de que se trata de uma alteração no ciclo da
14	1	em cerca de metade dos estudos . Ligeiramente menos relevantes ,	identificou -se	as alterações específicas na lubrificação vaginal (43 %) e as
15	1	Quanto aos fatores relacionados com a disfunção sexual durante a gravidez ,	verificou -se	que são diversificados (Tabela_2) . Da análise , é possível

Fonte: extraída do programa CQPweb.

Figura 3 – Linhas de concordância para a busca (5b).

Your query "[(word = "foi") | (word = "foram")]{1,3} [(pos = "VERB.Part.*")]" returned 95,129 matches in 8,799 different texts (in 48,506,519 words [9,900 texts]; frequency: 1,961.16 instances per million words) [2.523 seconds]

|< << >> >| Show Page: 1 Line View Show in random order New query Go!

No.	Text	Solution 1 to 50 Page 1 / 1903		
1	1	nos idiomas português , francês , espanhol e inglês . A pesquisa	foi realizada	nas bases de dados da EBSCHOhost (CINAHL , MEDLINE , Nursing
2	1	estudos na revisão . Por se tratar de uma revisão integrativa ,	foi realizada	pesquisa adicional nas referências dos artigos previamente selecionados e todos os resulta
3	1	pesquisa adicional nas referências dos artigos previamente selecionados e todos os resultados	foram integrados	, de modo a ter uma revisão mais ampla . A inclusão
4	1	modo a ter uma revisão mais ampla . A inclusão dos estudos	foi realizada	atendendo à concordância de dois investigadores de modo independente . Os dados
5	1	concordância de dois investigadores de modo independente . Os dados dos estudos	foram analisados	após preenchimento de um instrumento de colheita de dados com dados acerca
6	1	se encontravam em duplicado . Resultaram 68 estudos , dos quais 7	foram retirados	porque não cumpriam o critério de inclusão referente ao idioma e 3
7	1	47 estudos com texto completo (81 %) e os restantes	foram analisados	apenas a partir do resumo , uma vez que faziam parte de
8	1	uma vez que cerca de 40 % da produção sobre este tema	foi publicada	nos últimos 5 anos . Quanto à distribuição dos estudos neste período
9	1	anos . Quanto à distribuição dos estudos neste período , 24,1 %	foram publicados	em 2012 , seguindo -se 2014 com 22,4 % , 2011 e
10	1	com 5,1 % e Espanha e Israel com 3,5 % . Ainda	foram identificados	estudos oriundos de países como Canadá , Polônia , Austrália , Croácia
11	1	e na autoimagem (30 %) . Um pouco menos prevalentes	foram identificados	fatores religiosos , alterações na autoestima , presença de algumas vulnerabilidades (
12	2	aceroleira adubadas com fósforo e zinco . MATERIAL E MÉTODOS O experimento	foi instalado	em casa de vegetação situada no pomar do Departamento de Agricultura da
13	2	localizado na região . Após a retirada dos frutos , as sementes	foram lavadas	e semeadas em canteiros com areia lavada e , posteriormente , repicadas
14	2	cálcio e de magnésio p.a. , na relação 4:1 . A calagem	foi efetuada	20 dias antes da repicagem das mudas . O delineamento utilizado foi
15	2	misturado ao solo , até completa homogeneização . O sulfato de zinco	foi dissolvido	na água e aplicado via solução . Os demais nutrientes a as

Fonte: extraída do programa CQPweb.

O último caso em (5c) (Fig. 4) procura por ocorrências de verbos na forma finita na primeira pessoa do plural que não sejam precedidos por um pronome, por um substantivo ou pelo token "que".

Figura 4 – Linhas de concordância para a busca (5c).

Your query "[(pos != "PRON.*") & (pos != "NOUN.*") & (word != "que")]{1,3} [(pos = "VERB.Fin.Plur") & (word = ".*mos")]" returned 49,810 matches in 5,859 different texts (in 48,506,519 words [9,900 texts]; frequency: 1,026.87 instances per million words) [44.626 seconds]

|< << >> >| Show Page: 1 Line View Show in random order Choose action... Go!

No.	Text	Solution 1 to 50 Page 1 / 997		
1	1	e sem qualquer limite temporal , identificou -se 671 artigos , que	quando delimitamos	aos últimos 5 anos ficaram reduzidos a 266 artigos . Este fato
2	1	(3,5 %) artigos de opinião e dois (3,5 %	) editoriais . Destacamos	que 45 (77,5 %) são estudos originais , um (
3	1	) . Da análise comparativa com dos indicadores clínicos já existentes na	NANDA- I , verificamos	que cinco características definidoras já estão classificadas (alteração na atividade
4	3	como esta vinha se determinando até o presente momento . Por isso	, apresentaremos	, primeiramente , a constituição da liberdade de imprensa e o processo
5	3	da liberdade de imprensa e o processo de regulamentação em nível ocidental	. Depois , descreveremos	a constituição histórica da liberdade de imprensa no Brasil e o contraste
6	3	determinação arbitrária do editor do veículo.11 1.3 Liberdade de imprensa no Brasil	No Brasil , sabemos	que a Imprensa Nacional , órgão criado pelo Decreto de 13.05.1808 (
7	3	legislação específica , ou seja , um laissez-faire na atividade da imprensa	. Vejamos	como se articulam estas posições . 1.4 Lei de Imprensa ou ausência
8	3	no entanto , divergem quando se trata de encontrar uma solução	. Temos	duas posições : a) Necessidade de uma Lei de Imprensa :
9	3	da Filosofia do Direito . Porém , o que , de fato	, sabemos	é que o seu Prefácio expõe local e data , a saber
10	3	ou " caricatura insolente " do governo e de seus ministros .	Entretanto , não queremos	agora julgá -lo pelos padrões de hoje , e sim comparar suas
11	3	em algumas ocasiões , era até pretexto para perseguições e penalidades .	Além disso , sabemos	que Hegel vivenciou a experiência de censura , em 1808 , quando
12	3	imprensa e opinião pública a) Redator-chefe da Gazeta de Bamberg Hegel	, como já afirmamos	, trabalhou como diretor da Gazeta de Bamberg , de 1807 a
13	3	protagonistas , mostrar o que sabem fazer e expressar a sua opinião	. Temos	o MySpace , o YouTube , os blogs , lista de e-mails
14	4	de Pediatria da zona Norte do país e especialistas dedicados à Medicina	da Criança . Temos	ainda a colaboração de editores associados pertencentes a diversas áreas : Pediatria
15	4	pouco reconhecida , já que é habitualmente anônima . Neste último número	de 2012 agradecemos	, de modo particular , a todos os nossos revisores que contribuíram

Fonte: extraída do programa CQPweb.

Outras funcionalidades além das buscas por linhas de concordâncias são igualmente fáceis de se obter. Por exemplo, podemos identificar (sequências de) palavras que frequentemente coocorrem em um contexto, também conhecidos como

colocados. No CQPweb, podemos utilizar diferentes medidas estatísticas (ex.: MI, MI3, Z-score, T-score, log-likelihood, Coeficiente de Sorensen-Dice) para calcular e gerar uma lista de colocados. Por exemplo, a Fig. 5 mostra os 15 colocados mais fortes para a busca *(foram|foi)*, utilizando a medida estatística Log ratio (filtrada) e considerando apenas *tokens* ocorrendo à direita do nódulo e com a etiqueta gramatical de verbo no particípio. É importante notar que o etiquetador não tem 100% de precisão e algumas ocorrências (itens 2, 3 e 9, por exemplo) não estão no particípio.

Figura 5 – Lista de colocados para a busca (foram|foi).

Collocation based on:	Word form	Statistic:	Log Ratio (filtered)
Collocation window *from*:	1 to the Right	Collocation window *to*:	1 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: VERB.Part	VERB.Part — Submit changed parameters — Go!

Extra information:

The **Log Ratio** statistic is a measurement of *how big the difference is* between the (relative) frequency of the collocate alongside the node, and its (relative) frequency in the rest of the corpus or subcorpus.

In **the current collocation analysis**, all collocates displayed have Log-likelihood of at least **14.57104**.

The use of a log-likelihood filter means that it is not necessary to set high minimum values for *Freq(node, collocate)* and *Freq(collocate)* when using Log Ratio.

There are 77,321 different words in the collocation database for this query (Query "(foram|foi)" returned 171,230 matches in 9,361 different texts) [5.435 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log Ratio (filtered)
1	inaugurado	144	0.508	32	29	6.334
2	institucionalizando	25	0.088	5	5	6.141
3	evoluindo	159	0.561	29	28	5.977
4	decidido	334	1.179	56	50	5.829
5	verificado	1,605	5.666	263	218	5.79
6	desdobrado	32	0.113	5	5	5.708
7	admitido	215	0.759	33	31	5.678
8	excluído	300	1.059	46	43	5.676
9	surgindo	457	1.613	68	66	5.625
10	mantido	766	2.704	110	105	5.565
11	instalando	49	0.173	7	7	5.556
12	generalizando	36	0.127	5	4	5.509
13	revertido	52	0.184	7	7	5.456
14	respondido	204	0.72	25	24	5.301
15	forjando	41	0.145	5	5	5.293

Fonte: extraída do programa CQPweb.

Uma outra funcionalidade extremamente útil é a Distribuição. Com ela podemos ver, de forma clara, como os termos buscados se encontram distribuídos no *corpus* de acordo com os metadados. Tomamos como exemplo a busca descrita em (6).

(6)
[(word = "foi") | (word = "foram")] [(word = "observad.*") | (word = "verificad.*") | (word = "analisad.*")]

A primeira informação que obtemos é que, em todo o CoPEP, temos 3997 ocorrências do verbo "foi" ou "foram" seguido dos particípios passados de "observar", "verificar" e "analisar" em 2204 textos[13]. Com a função Distribuição, essa informação é calculada também para cada subseção do *corpus* (Tabelas 3, 4 e 5).

Tabela 3 – Distribuição de acordo com a variedade.

variedade	*tokens* na categoria	ocorrências na categoria	em n textos	de um total de n textos	Frequência por milhão de tokens na categoria
Br	24277363	2340	1212	3811	96,39
Eu	24229156	1657	992	6089	69,39

Fonte: elaborada pelas autoras.

Tabela 4 – Distribuição de acordo com a área de conhecimento.

área de conhecimento	*tokens* na categoria	ocorrências na categoria	em n textos	de um total de n textos	frequência por milhão de tokens na categoria
Ciências Agrícolas	2584817	946	442	907	365,98
Ciências Sociais Aplicadas	5376403	318	176	836	59,15
Engenharia	859758	85	47	176	98,87
Ciências Exatas e da Terra	803946	153	63	185	190.31
Ciências da Saúde	13382959	1958	1097	3996	146,31
Ciências Humanas	25498636	537	379	3800	21,06

Fonte: elaborada pelas autoras.

Tabela 5 – Distribuição de acordo com o ano de publicação.

anos	*tokens* na categoria	ocorrências na categoria	em n textos	de um total de n textos	frequência por milhão de tokens na categoria

[13] A justificativa para a seleção dos verbos "observar", "verificar" e "analisar" encontra-se na seção 3 deste trabalho.

1992-1997	162940	8	5	24	357,57
1998-2002	3358045	489	232	702	744,23
2003-2007	9259883	881	488	1878	480,97
2008-2012	19741497	1312	752	3974	332,86
2013-2018	15965038	13107	727	3322	521,55

Fonte: elaborada pelas autoras.

Os valores das tabelas acima nos permitem fazer comparações na frequência do uso da construção passiva nas diferentes categorias. Por exemplo, a frequência desse tipo de construção é bem maior nas publicações das Ciências Agrícolas do que nas Ciências Humanas, como mostra a Tabela 4.

4.3 Testes estatísticos

4.3.1 Teste de significância

A frequência por si só pode ser enganosa. Por isso, após obtermos os valores para as frequências, verificamos se a diferença entre duas observações é realmente significativa. Para essa verificação, aplicamos um teste de significância. Um teste de significância nos diz qual a probabilidade de se obter o resultado verificado em nossa amostragem, considerando que as variáveis sejam independentes (hipótese nula).

Assim, tomamos o exemplo em (6), na seção 4.2. A frequência normalizada do uso da passiva com os verbos "observar", "analisar" e "verificar" é maior em textos escritos em português brasileiro (96,39) do que em português europeu (68,39). Para verificar a chance da hipótese de que artigos em PB usem essa construção mais frequentemente que artigos escritos em PE, devemos aplicar um teste de significância. O teste de significância nos diz a probabilidade de termos o resultado que obtivemos com a nossa amostra no CoPEP se tivéssemos a certeza de que a variedade do português não está relacionada ao uso dessa construção.

Se a probabilidade for baixa, rejeitamos a hipótese nula e consideramos a nossa hipótese. Essa probabilidade é expressa como valor-p e varia entre 0 (não há chance de se obter os resultados observados) e 1 (certeza absoluta de se obter os resultados observados). Os valores de corte para o valor-p são arbitrários, mas geralmente considera-se que valores-p abaixo de 0.05 sejam significativos.

Para o caso supracitado, obtivemos um valor-p igual a $5{,}038995 \cdot 10^{-27}$ ao aplicar o teste exato de Fisher. Assim, podemos rejeitar a hipótese nula de que a variedade do português não influencia o uso da construção passiva e considerar nossa hipótese de que há diferença de uso entre as variedades europeia e brasileira. Caso o valor-p fosse acima de 0.05, aceitaríamos a hipótese nula e diríamos que a diferença na frequência observada não é significativa.

Existem diferentes testes de significância, sendo os três testes mencionados na sequência frequentemente utilizados na Linguística de *Corpus*. Chi quadrado é um teste amplamente utilizado em vários campos. Porém, apesar do seu alto uso também na Linguística de *Corpus*, este teste possui falhas. Entre essas falhas podemos ressaltar a baixa precisão e a baixa confiabilidade quando trabalhamos com números muito pequenos. O log-likelihood ou G2 teste é um teste amplamente utilizado na Linguística de Corpus (cf. DUNNING, 1993). O teste exato de Fisher, como o próprio nome diz, calcula a probabilidade exata.

4.3.2 Tamanho do efeito

Os testes de significância nos dizem a probabilidade de duas variáveis ou mais serem relacionadas. Para calcular a força com a qual elas estão associadas, utilizamos medidas de tamanho do efeito, também conhecidas como medidas de associação. Algumas medidas comumente usadas na Linguística de *Corpus* são o Coeficiente Phi e o V de Cramér. Para ambas as medidas, os valores vão de 0 (nenhuma correlação entre as variáveis) e 1 (as variáveis são completamente ligadas). O V de Cramér é

utilizado em situações que envolvem mais de duas variáveis. A interpretação usual para esses valores é a de que para valores próximos a 0,1 temos uma correlação baixa; uma correlação média para valores próximos a 0,3 e uma alta correlação caso este valor seja igual ou superior a 0,5. Para o nosso exemplo em (6), obtivemos um coeficiente Phi 0,0015, o que nos aponta a uma baixa correlação.

Os testes que apresentamos neste trabalho foram realizados no ambiente estatístico R, com o auxílio das bibliotecas *stats* (R CORE TEAM, 2018) e *rcompanion* (MANGIAFICO, 2016). Porém, há vários sites que oferecem serviço gratuito e simples de cálculo de testes estatísticos, como é o caso do Social Science Statistics[14].

5 Perspectivas e considerações finais

O presente trabalho apresentou recursos computacionais que permitem ampliar estudos sobre o fenômeno do desfocamento do agente, como os discutidos na seção três. O primeiro passo foi iniciar a preparação do *corpus* para estudar especificamente o contínuo de desfocamento do agente, como proposta apresentada na Fig. 1, bem como realizar algumas análises preliminares. A continuidade do trabalho tratará de, sem se limitar a, dois pontos principais.

Primeiro, faremos a identificação de cada seção (ex.: introdução, conclusão etc.) dos artigos presentes no *corpus*. Essa marcação nos permitirá restringir as buscas no *corpus* por seções do texto. Isso é particularmente útil no estudo do desfocamento do agente pois poderemos excluir seções relativas à revisão da literatura, por exemplo, concentrando-nos nas seções em que verbos como "observar", "analisar" e "verificar" tenham como agente os autores dos artigos, o que nos auxiliaria a localizar com mais precisão o que Morais (2013) identifica como baixo grau de desfocamento de participantes.

[14] Disponível em: https://www.socscistatistics.com/tests

O segundo ponto é implementar uma interface para realizar os testes de significância e tamanho do efeito diretamente na função distribuição do CQPweb. Um dos objetivos com essa nova funcionalidade é facilitar a validação das comparações feitas ao analisar o *corpus*.

Espera-se que o prosseguimento deste trabalho nos permita observar o contínuo de desfocamento do agente e testar a validade das propostas apresentadas neste trabalho, através da investigação de um *corpus* bem preparado através de uma ferramenta eficiente e confiável.

Esperamos também que os recursos aqui apresentados, seja o próprio *corpus* (CoPEP), seja as ferramentas utilizadas para anotação automática dos textos (Spacy), disponibilização e exploração do *corpus* (CQPweb), e aplicação de testes estatísticos (bibliotecas para R), possam servir como uma modesta exposição dos muitos recursos de que dispomos gratuitamente para pesquisas linguísticas.

Referências Bibliográficas

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri/SP: Manole, 2004.

BERBER SARDINHA, T.; PINTO, M. V. **Multi-dimensional analysis, 25 years on a tribute to Douglas Biber**. Amsterdam/São Paulo: John Benjamins Publishing Company/Universidade Católica de São Paulo, 2014. (Studies in Corpus Linguistics, ISSN 1388-0373; v. 60). DOI https://doi.org/10.1075/scl.60

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus linguistics**: investigating language structure and use. Cambridge: Cambridge University Press, 1998. DOI https://doi.org/10.1017/CBO9780511804489

CAMACHO, R. G. Construções de voz. *In*: ABAURRE, M. B. M.; RODRIGUES, A. C. S. (org.). **Gramática do português falado**. Novos estudos descritivos. Campinas: UNICAMP, 2002. p. 227-316.

DUNNING, T. E. Accurate methods for the statistics of surprise and coincidence. **Computational linguistics**, v. 19, n. 1, p. 61-74, 1993.

HARDIE, A. CQPweb — combining power, flexibility and usability in a corpus analysis tool. **International Journal of Corpus Linguistics**, [S.L.], v. 17, n. 3, p. 380-409, 31 dez. 2012. DOI https://doi.org/10.1075/ijcl.17.3.04har

KILGARRIFF, A.; BAISA, V.; BUŁTA, J.; JAKUBÍčEK, M.; KOVÁř, V.; MICHELFEIT, J.; RYCHLÝ, P.; SUCHOMEL, V. The Sketch Engine: ten years on. **Lexicography**, [S.L.], v. 1, n. 1, p. 7-36, jul. 2014. DOI https://doi.org/10.1007/s40607-014-0009-9

KUHN, T. Z.; FERREIRA, J. P. O Corpus de Português Escrito em Periódicos - CoPEP. **Delta**: Documentação de Estudos em Lingüística Teórica e Aplicada, [S.L.], v. 36, n. 2, p. 1-42, fev. 2020. DOI https://doi.org/10.1590/1678-460x2020360209

LEECH, G. Corpora and theories of linguistic performance. *In*: **Directions in corpus Linguistics**. Proceedings of Nobel Symposium 82. Stockolm, 4-8 Aug 1991. Dan Svartvik (editor). Morvton de Guryter: Berlin, 1992. p.105-122.

LEECH, G. Review of Biber, Conrad, and Reppen. Corpus linguistics: Investigating language structure and use. **International Journal of Corpus Linguistics**, Amsterdã, John Benjamins, 4(1), p. 185-88, jun/1999. DOI https://doi.org/10.1075/ijcl.4.1.11lee

MANGIAFICO, S.S. **Summary and Analysis of Extension Program Evaluation in R**, version 1.18.8, 2016. Disponível em: https://rcompanion.org/handbook. Acesso em: 1 out. 2021.

MORAIS, F. B. C. de. **Entre alhos e bugalhos - os usos do clítico *se* na escrita acadêmica**. 2013. 183 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2013.

MORAIS, F. B. C. de. Variação de usos do clítico *se* na comunidade acadêmica: um estudo descritivo com base na linguística sistêmico-funcional. **Cadernos de Linguagem e Sociedade**, v. 17, n. 1, p. 70-100, 2016. Disponível em: https://periodicos.unb.br/index.php/les/article/view/4429. Acesso em: 22 jul. 2021. DOI https://doi.org/10.26512/les.v17i1.4429

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2018. Disponível em: https://www.R-project.org. Acesso em: 1 out. 2021.

ROCHA, M. A. E. Introduction to the issue on corpus linguistics. **Ilha do Desterro**, n. 52, Florianópolis, p. 9-33, jan./jun. 2007.

SCHMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. **Proceedings of International Conference on New Methods in Language Processing**. Manchester, UK.: [s.n.]. 1994.

SHIBATANI, M. Passives and related constructions: a prototype analysis. **Language**, v. 61, n. 4, p. 821-848, 1985. Disponível em: https://www.researchgate.net/publication/244437975_Passives_and_Related_Constructions_A_Prototype_Analysis. Acesso em: 9 mar. 2021. DOI https://doi.org/10.2307/414491

SINCLAIR, J. McH. From theory to practice. *In*: LEECH, G.; MYERS, G., THOMAS J. **Spoken English on computer**: transcription, mark-up and application. Londres: Longman, 1995. p 99-112.

Artigo recebido em: 15.12.2021 Artigo aprovado em: 21.02.2022