

Contribuições da análise colexêmica distintiva para o estudo da construcionalização

Contributions of distinctive collexeme analysis to the study of constructionalization

Monclar Guimarães Lopes*
Karen Sampaio Braga Alonso**

RESUMO

A Análise Colexêmica Distintiva (ACD) é um método quantitativo que compara construções similares quanto às preferências colocacionais. Assim como Hilpert (2006), defendemos que a ACD também seja útil para mapear mudanças ao longo do tempo, como no caso em que as relações sequenciais são reconfiguradas em decorrência de processos de automatização. Neste estudo, partimos da hipótese de que existe um marcador estruturante do discurso (MED) com valor de acréscimo cujas partes constituintes principais são a preposição *sem* seguida de verbo *dicendi*, que seria o resultado da recategorização de uma oração hipotática adverbial (OHA) modal ou condicional formada por esses mesmos elementos. Ao todo, analisamos 754 ocorrências da sequência *sem* + verbo *dicendi* no Portuguese Corpus 2022. Identificamos que as duas construções apresentam diferenças quanto à possibilidade de preenchimento de *slot*: 22 verbos *dicendi* aparecem na construção de OHA, ao passo que apenas 6 ocorrem na construção de MED, sendo dois deles (*contar* e *falar*) altamente atraídos para esta última construção.

Palavras-chave: Análise colexêmica distintiva. Construcionalização. Relações Sequenciais. Chunking. Linguística Funcional Centrada no Uso.

Recebido em 2 de junho de 2025.

Aceito em 30 de agosto de 2025.

DOI: <https://doi.org/10.18364/rc.2026n70.1484>

* Universidade Federal Fluminense, monclarlopes@id.uff.br
<http://orcid.org/0000-0002-6238-958X>

** Universidade Federal do Rio de Janeiro, karensampaio@letras.ufrj.br
<https://orcid.org/0000-0002-7853-0015>

ABSTRACT

Distinctive Collexeme Analysis (DCA) is a quantitative method that compares similar constructions regarding their collocational preferences. Like Hilpert (2006), we argue that the DCA is also useful for mapping changes over time, such as when sequential relations are reconfigured due to processes of automatization. In this study, we hypothesize that there exists a discourse structuring marker (DSM) with an additive value, whose main components are the preposition “sem” followed by a speech verb, which would result from the recategorization of a modal or conditional adverbial hypothetical clause (AHC) formed by the same elements. In total, we analyzed 754 tokens of the sequence “sem” + speech verb in the Portuguese Corpus 2022. We identified that the two constructions differ in their instantiation of the speech verb slot: 22 speech verbs appear in AHC, while only 6 occur in the DSM construction, with two of them (“contar” and “falar”) being highly attracted to the latter.

KEYWORDS: Distinctive Collexeme Analysis. Constructionalization. Sequential Relations. Chunking. Usage-Based Linguistics.

Introdução

Os Modelos Baseados no Uso (MBU) oferecem um conjunto de evidências que sugerem que há uma relação estreita entre o uso da língua e a forma como a representamos na nossa mente. Tendo isso em vista, métodos qualiquantitativos vêm sendo aplicados a dados empiricamente atestados, que visam, de um lado, interpretar tais dados qualitativamente, em termos de como se manifestam na língua em uso e impactam o conhecimento linguístico do falante. De outro lado, em termos quantitativos, demonstrar que a forma como representamos cognitivamente o conhecimento linguístico também é impactado pela distribuição estatística e a frequência com que determinadas estruturas ocorrem na nossa experiência com a língua.

O aspecto quantitativo também está associado à noção de produtividade, no que diz respeito às frequências *type* e *token*. Segundo

Bybee (2003), a frequência *token* está associada à extensão de uso de um determinado construto por parte dos falantes, ao passo que a frequência *type*, à expansão *host-class* (cf. Himmelmann, 2004). Nesse último caso, as construções, à medida que se tornam mais esquemáticas, passam a possibilitar novas colocações e, por isso, originam novas microconstruções de função análoga. O levantamento da produtividade é normalmente aferido por meio de frequência bruta e relativa, mas é possível (e desejável) agregar a essas medidas alguns cálculos estatísticos mais robustos, sobretudo quando lidamos com um grande conjunto de dados. Isso decorre do fato de os modelos matemáticos serem muito bons na busca de associações entre variáveis, que podem ser um tanto obscuras quando interpretadas a olho nu.

Nesse sentido, há um conjunto de testes estatísticos particularmente desenvolvidos para a abordagem construcional, inicialmente elaborados por Stefanowitsch e Gries (2003, 2004), abrigados sob o rótulo *abordagem colostrucional*. Neste texto, exploraremos um deles: a análise colexêmica distintiva (ACD). A ACD busca contrastar duas ou mais construções, funcionalmente relacionadas ou não, quanto às suas preferências colocacionais. Embora o método tenha sido desenvolvido, a priori, para a descrição de construções em perspectiva sincrônica, Hilpert (2006) propõe que ele também seja útil para mapear diferenças de produtividade de uma construção ao longo do tempo, uma vez que novos *types* podem surgir ou se tornar obsoletos.

Assim como Hilpert (2006), neste artigo, propomos que a ACD também seja uma boa ferramenta para tratar casos de construcionalização gramatical em construções complexas, cujas relações sequenciais (cf. Diessel, 2019) tenham sido reconfiguradas em razão da atuação de processos de automatização, especialmente aqueles que envolvem *chunking*. Como ilustração da potencialidade desse tipo de análise, apresentamos, neste texto, um estudo sobre a construção [D1 sem V_{dicendi} (que) D2]¹, que classificamos,

1 Neste texto, utilizaremos as seguintes siglas para as notações das construções: D1 e D2 representam os segmentos discursivos articulados por *sem Vdicendi (que)*, que podem ser

com base nas investigações de Traugott (2022), como um marcador estruturante do discurso (MED) com valor de acréscimo. Segundo Lopes (2025) e Lopes e Oliveira (2024), essa construção originou-se no século XVIII como resultado da recategorização da oração hipotática adverbial (OHA) modal e/ou condicional negativa formada, inicialmente, por *sem* + *contar*. Posteriormente, no século XX, essa construção, por meio de analogização, teria dado origem a novos *types*, constituídos por *sem* + verbo *dicendi*, como *falar*, *dizer*, *mencionar* etc.

Na pesquisa proposta para este texto, buscamos aferir, para além da extensibilidade da construção de MED no que diz respeito aos lexemas que se justapõem à preposição *sem*, o grau de associação desses mesmos lexemas tanto com a construção de MED quanto com a de OHA. Partimos da hipótese de que verbos muito fortemente associados a uma das duas construções tendem a ser pouco produtivos (ou até improdutivos) na outra, em decorrência da atuação dos processos relativos à memória, sobretudo no que diz respeito à economia de seu processamento. Afinal, se uma sequência de elementos é muito frequentemente empregada para um determinado uso, ela tende a tornar-se fortemente associada a ele, o que, por consequência, desfavorece seu uso em contextos distintos, via *bloqueio estatístico* (Goldberg, 2019).

A investigação foi realizada com base nos pressupostos da Gramática de Construções Baseada no Uso (cf. Goldberg, 1995; Hilpert, 2014; Diessel, 2019, entre outros) e os dados analisados foram extraídos do *Portuguese Corpus 2022*, do Wortschatz Leipzig², constituído exclusivamente de notícias eletrônicas em português. Ao todo, analisaram-se 754 ocorrências constituídas da preposição *sem* justaposta a um conjunto de 22 verbos *dicendi*.

No intuito de atender à estrutura de trabalho proposta, este texto está estruturado em cinco seções. Além desta introdução, há: pressupostos

de diferentes níveis: suboracionais, oracionais e supraoracionais; *sem* é a parte fixa da construção; *Vdicendi* é um slot construcional que prevê a instanciação de verbos *dicendi*; (*que*) é representada entre parênteses porque é uma subparte facultativa, podendo ou não ocorrer na construção.

2 Disponível em: <https://wortschatz.uni-leipzig.de/de> - acesso em 20 mar. 2025.

teóricos; estudo de caso e metodologia; resultado e discussão; considerações finais. Por fim, fecha-se o texto com as referências.

1. Pressupostos teóricos

No âmbito dos MBU, a Gramática de Construções Baseada no Uso (GCBU) é um modelo de descrição gramatical que prevê que o conhecimento linguístico dos falantes pode ser representado por uma rede de construções linguísticas. Nessa abordagem, as construções – definidas como pareamentos de forma e significado (Goldberg, 1995; Croft, 2001, Hilpert, 2014, Diessel, 2019, entre outros) ou unidades simbólicas convencionais (Langacker, 1987) – “são generalizações sobre a experiência concreta, representadas mentalmente como o conhecimento que um indivíduo tem de sua língua” (Lopes e Oliveira, 2024, p. 195).

A rede construcional – ou *constructicon* –, é compreendida por meio da metáfora da rede, a partir da qual haveria um conjunto de nós (ou construções) e de relações entre nós (ou link construcionais). Segundo Diessel (2019), as construções distribuem-se no *constructicon* em relações de diferentes níveis: simbólicas, sequenciais, taxonômicas, lexicais, relações horizontais entre construções e de preenchimento de *slot*.

As relações taxonômicas – que preveem representações construcionais em diferentes níveis de esquematicidade – são especialmente caras à descrição linguística. Dentre outros aspectos, elas nos permitem mapear a produtividade das construções linguísticas em termos de sua frequência *type* ou *token*. A formação de (novos) *types* costuma ser explicada com base no princípio da coerência semântica (cf. Goldberg, 1995, p. 50), para o qual “apenas papéis que sejam semanticamente compatíveis podem se fundir”³. Como ilustração, vejamos duas ocorrências da construção de movimento causado, extraídos de Goldberg (1995, p. 87; 55):

3 No original: “Only roles which are semantically compatible can be fused”.

- (a) Joe moved it onto the table (Joe o colocou sobre a mesa)
- (b) Fred sneezed the napkin off the table (Fred espirrou o guardanapo para fora da mesa)

Segundo a autora (Goldberg, 1995), a construção de movimento causado pode ser representada pela estrutura argumental [X CAUSA Y a MOVER Z], em que um sujeito agente (X) é o causador de uma mudança de localização (Z) de um objeto (Y). Esse tipo de estrutura prevê a instanciação de um tipo semântico de verbo de movimento com três argumentos, que implica uma ação verbal em que um AGENTE desloca um PACIENTE para um determinado ALVO⁴, como é o caso de *to move* em (a). Não obstante esse fato, a construção é extensível a outros verbos que, em seus contextos de uso (prototípicos ou não), também permitiriam o deslocamento espacial de um objeto Y por meio da ação de um agente X, como podemos observar em (b). Nesse sentido, a despeito de *to sneeze* não ser um verbo originalmente mobilizado para desencadear movimento de objetos, ele é coerente com a estrutura em que é empregado. A mesma coerência não seria possível se o verbo fosse *to whisper* (suspirar), dado que essa ação não apresentaria força suficiente para promover uma mudança espacial de Y.

Embora o verbo *to sneeze* seja compatível com a construção de movimento causado, entendemos que ele representa um uso idiossincrático para essa construção, cujos verbos mais prototípicos são *mover*, *deslocar*, *colocar*, *pôr* e outros com propriedades semânticas análogas. Na visão da GCBU, os esquemas construcionais são o resultado da abstratização das propriedades de forma e significado dos lexemas que instanciam as construções. Sob esse ponto de vista, se as propriedades da construção de movimento causado advieram mais extensivamente de verbos como *mover*, *deslocar*, *colocar* e

4 Tomamos a liberdade de fazer uma adaptação dos papéis semânticos atribuídos por Goldberg (1995, p. 52), que associa ao sujeito o papel de *causa*; ao objeto, de *tema*; ao adjunto (SP), de *alvo*.

pôr, isso se deve especialmente à frequência com que esses elementos são instanciados por essa construção. Stefanowitsch e Gries (2004) chegam a esse mesmo tipo de conclusão no que diz respeito ao emprego do verbo *to give* nas construções ditransitiva e dativa preposicionada.

	<i>to give</i>	<i>outros verbos</i>	<i>Total</i>
Construção ditransitiva	461 (44,54%)	574 (55,46%)	1.035 (100%)
Construção dativa preposicionada	146 (7,60%)	1.773 (92,40%)	1.919 (100%)

Tabela 1. Distribuição de *give* na construção ditransitiva e na dativa preposicionada. Fonte: Stefanowitsch e Gries, 2004, p. 102 (adaptado).

Na Tabela 1, verificamos que o verbo *to give* é mais frequente na construção ditransitiva (461) do que na dativa preposicionada (146). Paralelamente, observamos que, das 1.035 ocorrências analisadas da construção ditransitiva, *to give*, sozinho, representa 44,54% dos usos. Ou seja, quase metade dos usos da construção ditransitiva concorre com o verbo *to give*, o que sugere uma relação forte entre o sentido do verbo e o da construção.

Como podemos inferir, os dados da Tabela 1 foram extraídos com base em um único fator, qual seja, a frequência de um determinado lexema na construção. Bybee (2010, p. 97) defende que “a frequência de um lexema L na construção é o fator mais importante, e que talvez sua frequência relativa desempenhe um papel em relação à frequência total da construção”⁵. No entanto, esse ponto de vista não é compartilhado por todos os estudiosos. Stefanowitsch e Gries (2003), Hilpert (2006) e Stefanowitsch e Flach (2020), por exemplo, recomendam que se comparem quatro frequências: (1) a frequência de ocorrência de um determinado lexema em todo o *corpus*; (2) a frequência de ocorrência desse mesmo lexema na construção-alvo; (3) a frequência de ocorrência de outros lexemas (que concorram pelo uso com

5 No original: “the frequency of the lexeme L in the construction is the most important factor with perhaps the frequency relative to the overall frequency of the construction playing a role”.

o lexema investigado) na construção-alvo; (4) a frequência de ocorrência de todas as construções no *corpus*. Essas medidas possibilitam ao analista, por meio da utilização de métodos computacionais estatísticos, determinar o quanto um lexema é esperado ou não dentro da construção (em comparação com a frequência observada), bem como indicar sua força de associação com a construção, isto é, em que medida ele é atraído ou repellido.

Sob esse ponto de vista, retomando-se os dados expressos na Tabela 1, o novo cálculo levaria em consideração a frequência *type* e *token* de todos os lexemas (verbos) que instanciam as construções ditransitiva e dativa preposicionada, bem como a frequência desses lexemas em outras construções e a quantidade total de construções no *corpus*. Empregando-se as fórmulas estatísticas necessárias com base nesses valores, o resultado indicará, para cada um dos lexemas, qual é o seu grau de associação com as construções investigadas. A despeito de incompletos para esse tipo de análise, os dados da Tabela 1 nos permitem inferir que o grau de associação de *give* para a construção ditransitiva seja, muito provavelmente, relativamente alto, de modo que ele deve ocorrer nessa construção mais frequentemente do que seria previsto por puro acaso (cf. Stefanowitsch et al., 2005, p. 646). Trata-se de uma dedução correta, como podemos verificar na Tabela 2.

Construção ditransitiva (N=1.035)		Construção dativa preposicionada (N= 1.919)	
Colexemas	distintividade	Colexemas	distintividade
give (461:146)	1.84E-120	bring (7:82)	1.47E-09
tell (128:2)	8.77E-58	play (1:37)	1.46E-06
show (49:15)	8.32E-12	take (12:63)	0.0002
offer (43:15)	9.95E-10	pass (2:29)	0.0002
cost (20:1)	9.71E-09	make (3:23)	0.0068
(...)	(...)	(...)	(...)

Tabela 2. Colexemas distintivos entre as construções ditransitiva e dativa preposicionada.

Fonte: Stefanowitsch & Gries, 2004, p. 106 (adaptado).

Na Tabela 2, a primeira coluna traz os lexemas identificados mais recorrentemente em cada uma das construções, com sua indicação numérica entre parênteses. Desse modo, para *to give*, há 461 ocorrências na construção ditransitiva e 146 para a dativa preposicionada; para *to bring*, há 7 na ditransitiva e 82 na dativa preposicionada; e assim sucessivamente. A segunda coluna, por sua vez, apresenta a medida de distintividade, calculada com base no valor exato de Fisher (isto é, a probabilidade estatística de que haja erro entre a associação dos lexemas com as respectivas construções). No caso de *give*, a notação científica 1.84E-120 significa que as chances de erro na associação (ou seja, de aleatoriedade na seleção) desse verbo para a construção ditransitiva é menor do que 8.77 multiplicado por 10 elevado a -120. Isso representa 120 casas decimais depois da vírgula, ou seja, um número extremamente pequeno. Em outras palavras, quanto mais baixo for esse número, maior deve ser o grau de associação entre o lexema e a construção⁶. E, como podemos notar, *give* representa o número mais baixo de todos os dados da Tabela 2.

Esse método de análise, intitulado *análise colostrucional*⁷, fora desenvolvido por Stefanowitsch e Gries (2003) e se destina a três tipos distintos de análise: (1) a *análise colexêmica simples*, que busca avaliar o grau de associação de um ou mais lexemas em uma construção específica; (2) a *análise colexêmica distintiva*, que busca ver o grau de associação de um ou mais lexemas em duas ou mais construções; (3) *análise colexêmica covariacional*, que busca analisar o comportamento de dois ou mais lexemas em uma única construção. Em virtude do escopo de nosso texto, iremos explorar apenas os aspectos atinentes à *análise colexêmica distintiva* (ACD).

A ACD envolve, preferencialmente, a análise de um par alternante de construções. Busca “identificar lexemas que apresentam uma forte

6 É importante mencionar que, nos cálculos estatísticos, o valor de p refere-se ao teste da hipótese nula (H0), o qual busca verificar se a associação entre duas ou mais variáveis pode ser entendida como aleatória. Sendo assim, quanto menor for o valor de p, menores são as chances da inexistência de associação (aleatoriedade) entre as variáveis comparadas.

7 Do inglês, *collostructional analysis*, uma composição dos termos *collocations* + *constructions*.

preferência por um membro do par em detrimento do outro, e assim torna possível identificar diferenças distributivas sutis entre os membros de tal par”⁸ (Gries; Stefanowitsch, 2004, p. 97). Nas palavras dos autores (2004, p. 101): “nós propomos um método (...) em que olham para construções quase sinônimas (ou funcionalmente quase equivalentes) (...) e focalizam as palavras que aparecem recorrentemente em seus *slots*”⁹. Às palavras que coocorrem no mesmo *slot* construcional, os autores chamam de *colexemas* da construção. Sob esse ponto de vista, os dados da Tabela 1 se prestam intimamente à ACD, sendo o par alternante a construção ditransitiva e a dativa preposicionada, que compartilham entre si a ação de transferência de um PACIENTE entre um AGENTE e um BENEFICIÁRIO – e.g.: *I gave him a book* vs. *I gave a book to him* –, e os colexemas que concorrem para o *slot* do verbo, formas como *give* e *bring*, por exemplo – e.g.: *I brought him a book* vs. *I brought a book to him*.

Como mencionamos previamente, Hilpert (2006) argumenta que esse mesmo método pode ser valioso para o estudo de fenômenos de mudança, resultantes de processos de gramaticalização¹⁰. Isso decorre do fato de haver uma relação direta entre a mudança nas preferências colocacionais da construção ao longo do tempo e a mudança semântica. Essa comparação, quando feita exclusivamente com base em frequência bruta, segundo o estudioso, pode ser um tanto mais difícil ao analista, uma vez que, na investigação diacrônica, tende-se a trabalhar com um grande conjunto de dados. Além disso, o olho nu do analista nem sempre é capaz de perceber diferenças sutis, que podem ser capturadas com o auxílio de ferramenta estatística.

8 No original: “(...) identifies lexemes that exhibit a strong preference for one member of the pair as opposed to the other, and thus make it possible to identify subtle distributional differences between the members of such a pair”.

9 No original: “We propose a method (...) in that we look at near-synonymous (or functionally near-equivalent) constructions (...) that we focus on words appearing in particular slots in these constructions”.

10 Neste texto, para falar de fenômenos de mudança que afetam o estatuto gramatical dos usos, empregaremos o termo *construcionalização*, e não *gramaticalização*, conforme proposta de Traugott e Trousdale (2013).

Como ilustração, Hilpert (2006) apresenta um novo estudo sobre o modal *shall* com base nas investigações prévias de Traugott (1989), que observou que, desde o século XVI, esse verbo tem passado por mudança semântica, em que o sentido de obrigação é recategorizado para o de futuridade. De acordo com a estudiosa (1989), esse novo significado está associado a uma trajetória em direção a significados que estão cada vez mais situados nas crenças ou atitudes do falante – fenômeno conhecido como subjetificação (Traugott; Dasher, 2002). Segundo Hilpert (2006), “as descobertas de Traugott preveem que os usos mais recentes de *shall* **devem** mostrar uma afinidade mais forte com complementos verbais que expressam significados subjetivados, baseados no falante” (Hilpert, 2006, p. 252, *grifo nosso*)¹¹. Seguindo essa linha de raciocínio, Hilpert (2006) busca verificar a força colostrucional dos lexemas que se justapõem a *shall*. Seus resultados corroboram as hipóteses de Traugott (1989), na medida em que verbos de natureza mais subjetiva começam a apresentar um grau de associação relativamente alto com *shall* a partir do século XVI, como *understand*, *forget*, *suppose*, *think* e *feel*, por exemplo, conforme é possível observar na Tabela 3.

1500-1640			1640-1780			1780-1920		
Verbo	N	FColost	Verbo	N	FColost	Verbo	N	FColost
<i>understand</i>	48	15.48	<i>endeavour</i>	52	16.36	<i>forget</i>	81	17.01
<i>come</i>	120	10.32	<i>discover</i>	17	7.86	<i>go</i>	194	12.91
<i>forfeit</i>	40	6.53	<i>examine</i>	13	6.86	<i>get</i>	81	9.46
<i>perceive</i>	19	6.52	<i>mention</i>	18	5.90	<i>try</i>	27	6.87
<i>bear</i>	30	6.49	<i>suppose</i>	14	5.67	<i>meet</i>	53	6.36
<i>appear</i>	37	5.65	<i>confine</i>	10	5.29	<i>feel</i>	32	5.59
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)

Tabela 3. 15 colexemas mais distintivos de *shall* em três períodos de tempo. Fonte: Hilpert, 2006, p. 249. (adaptado).

11 No original: “Traugott’s findings predicts that later uses of *shall* should show a stronger affinity to verbal complements that express subjectified, speaker-based meanings”.

Neste artigo, argumentamos que a ACD também possa ser interpretada como um método adequado para outro tipo de investigação da mudança linguística: aquela que envolve duas construções distintas (A e B), sendo a segunda o resultado da construcionalização gramatical da primeira. Defendemos que essa abordagem seja particularmente interessante para a análise de descrições complexas, em que as relações sequenciais (cf. Diessel, 2019) da construção B tenham sofrido processos de automatização, em especial *chunking*, e consequente perda de composicionalidade, seja semântica e/ou sintática.

Conforme Diessel (2019, p. 63),

a língua é um meio linear em que todos os elementos linguísticos – sons da fala, morfemas, palavras, sintagmas, cláusulas e sentenças – estão organizados em ordem sequencial. (...) Muito frequentemente as sequências linguísticas são fixas (e convencionalizadas) como o resultado da automatização (e da cognição social). Se a mesma sequência de elementos linguísticos é processada repetidamente, a automatização cria *links* sequenciais entre eles¹².

O resultado direto da automatização é a formação de *chunks*, isto é, de sequências de elementos que são processadas como uma unidade, um bloco único de forma e sentido. Segundo Bybee (2010, p. 36),

O estatuto de um *chunk* na memória distribui-se num *continuum*. Certamente palavras que nunca foram usadas juntas não constituem um *chunk*, mas, por outro lado, existe um *continuum* que vai de palavras que foram utilizadas juntas somente uma vez e bem recentemente, as quais vão constituir um *chunk* fraco, cujas partes internas são mais fortes do que o todo, até *chunks* mais frequentes, como *lend a hand* (“dar uma mão”) e *pick and choose*

12 No original: “Language is a linear medium in which all linguistic elements – speech sounds, morphemes, words, phrases, clauses and sentences – are arranged in sequential order. (...) Very often linguistic sequences are fixed (and conventionalized) as a result of automatization (and social cognition). If the same string of linguistic elements is repeatedly processed, automatization creates sequential links between them”.

(“escolher a dedo”), que, são mais facilmente acessíveis como um todo, embora ainda mantenham ligação com suas partes¹³.

A despeito de haver *chunks* de diferentes graus de vinculação, alguns constituídos de formas bastante fixas e rígidas – como alguns tipos de idiomatismos (e.g.: *puxa-saco*) –, e outros mais esquemáticos, que permitem a alternância de algumas subpartes (e.g.: *em prol/favor de*), uma das consequências do *chunking* é a redução nas possibilidades de escolha. Ou seja, quando comparamos uma construção A (forma-fonte, menos procedural) com uma construção B (forma-nova, mais procedural), espera-se que esta última apresente um paradigma de seleção de lexemas um tanto menor em relação àquela. Além disso, ainda sobre esse aspecto, também se espera que, à medida que a construção B se torna mais produtiva, seus lexemas mais frequentes tornem-se cada vez mais associados a ela e, ao mesmo tempo, menos associados à construção A, em decorrência das ações cognitivas que priorizam a economia do processamento da memória.

Para ilustrar a potencialidade desse tipo de estudo, apresentamos, na seção seguinte, o emprego da ACD para a descrição do marcador estruturante do discurso (MED) de acréscimo [D1 sem V_{dicendi} (que) D2].

2. Estudo de caso e metodologia

Segundo a literatura gramatical e linguística, a preposição *sem* apresenta o sentido básico de subtração/negação, ausência e desacompanhamento (cf. Cunha e Cintra, 2001; Rocha Lima, 1972), conforme podemos identificar em usos como: a) *xícara sem asa*; b) *comida sem gosto*; c) *chegou sem o marido*. Quando ela introduz uma oração hipotática adverbial reduzida, esses

13 No original: “the status of a chunk in memory falls along a continuum. Certainly, words that have never been experienced together do not constitute a chunk, but otherwise there is a continuum from words that have been experienced together only once and fairly recently, which will constitute a weak chunk whose internal parts are stronger than the whole, to more frequent chunks such as *lend a hand* or *pick and choose* which are easily accessible as wholes while still maintaining connections to their parts”.

mesmos valores semânticos são mobilizados para a expressão de noções de modo e condição negativos (cf. Neves, 2018; Raposo et al., 2015). A título de ilustração, seguem duas ocorrências:

(01) Um jovem de 23 anos morreu após ter sido atingido por um tiro efetuado acidentalmente pelo pai no Bairro de São Luís, em Divinópolis. Segundo a Polícia Militar (PM), a família contou que o disparo aconteceu enquanto ele limpava a arma. A família ainda disse que depois que se deu conta do ocorrido. Ele saiu *sem contar* onde iria. A polícia configura como fuga.

(02) Nesta terça-feira (11), em São Paulo, o amigo do jogador Neymar André Galo, responsável pela estadia da modelo Najila Trindade em Paris e pela compra de suas passagens, prestou depoimento. Ele saiu da delegacia *sem falar* com a imprensa.

Em (01) e (02), *sem contar onde iria* e *sem falar com a imprensa* são orações hipotáticas adverbiais modais (OHA) reduzidas de infinitivo, cujo escopo é o verbo *sair*, na oração matriz. Lopes (2025) e Lopes e Oliveira (2024) têm representado essas orações, em termos construcionais, da seguinte maneira: [OM_{D1} [[sem][verbo_{infinitivo} (X)]]_{D2}]. Os colchetes externos representam todo o período composto por hipotaxe, em que há dois segmentos discursivos – D1 e D2: um constituído pela oração matriz (OM); outro pela oração hipotática adverbial reduzida de infinitivo encabeçada por *sem*. Os colchetes internos, que separam *sem* do verbo, foram colocados para marcar uma diferença em relação às propriedades de uma construção que veremos logo a seguir. Nesta notação, os elementos são menos vinculados entre si, com um pouco mais de autonomia semântica. A preposição *sem* mantém seus sentidos básicos de negação e ausência, assim como os verbos mantêm suas propriedades semânticas mais prototípicas. Além disso, o verbo pertence a um paradigma relativamente amplo, na medida em que, a priori, quaisquer formas verbais podem concorrer para essa posição.

Além desses usos, já classificados e disseminados pela literatura gramatical e linguística, Lopes (2025), Lopes e Oliveira (2024) e Lopes e Moura (2022) flagram um outro no português contemporâneo, em que *sem*, adjungido a um verbo *dicendi*, como *contar*, *falar*, *dizer*, *mencionar* etc., recategoriza-se como um conector de acréscimo – aqui classificado como um *marcador estruturante do discurso* – ou MED – (cf. Traugott, 2022)¹⁴. Abaixo, apresentamos duas ocorrências como ilustração desses usos:

(03) Com a alta procura nos postos de saúde por pacientes com sintomas de Covid-19, os testes começaram a faltar, *sem contar que* a maioria dos postos precisou suspender consultas agendadas anteriormente.

(04) Daqueles episódios que provam porque *Friends* é considerada uma das maiores séries de comédia de todos os tempos. Quase tudo é impecável. Temos Chandler vestido de coelho rosa, Monica de Mulher-Gato, Phoebe de Supergirl, Joey de Chandler e Rachel tentando usar suas roupas boas antes de avançar na gravidez. *Sem falar* na inacreditável fantasia de Ross, facilmente confundida como um “cocô espacial”.

Em (03) e (04), *sem contar* e *sem falar*, apesar de apresentarem estrutura tipicamente hipotática – uma preposição seguida de verbo no infinitivo, característica típica de orações adverbiais reduzidas –, não escopam um verbo (ou algum elemento sintático) presente na oração matriz, atribuindo-lhe uma noção circunstancial de modo ou de condição. Na verdade, o escopo dessas estruturas é mais amplo, na medida em que engloba o enunciado prévio (que pode ser constituído de mais de uma oração, período

14 Os marcadores estruturantes do discurso – ou MEDs – representam uma categoria discursiva representada por diferentes classes gramaticais e expressões linguísticas, que atuam “como conectores que permitem ao falante/escritor sinalizar que tipo de relação ele deseja que o ouvinte/leitor deduza da ligação dos elementos do discurso” (Traugott, 2022, p.4).

ou até parágrafo). Paralelamente, a estrutura é mobilizada em um contexto em que é recategorizada para introduzir um argumento convergente, a título de acréscimo. Nesse sentido, há uma certa dessemantização da preposição *sem*, cujos sentidos de negação e ausência encontram-se esmaecidos para dar lugar à semântica da adição. Em (03), por exemplo, há dois fatos enumerados, em direção convergente, que demonstram a preocupação do enunciador: a falta dos testes para Covid-19 E a suspensão de consultas agendadas anteriormente na maioria dos postos. Igualmente, em (04), há várias vestes engraçadas utilizadas pelos personagens de Friends, **ALÉM DA** “inacreditável fantasia do Ross”.

Para além desses aspectos, pode-se inferir que, nessa construção, o falante faz uma espécie de gerenciamento de atenção intersubjetiva no discurso, não agindo diretamente sobre o conteúdo proposicional. Ele evocaria indiretamente um argumento novo formalmente apresentado como dispensável, mas que reforçaria, na verdade, a sua argumentação, na medida em que ativa o domínio conceptual referente ao conteúdo vinculado a esse argumento. Logo, entendemos que o falante recruta [D1 sem V_{dicendi} (que) D2] para seu discurso mais como um recurso para gerenciar a atenção do ouvinte, dando relevo ao argumento expresso em D2, do que para acrescentar apenas mais um argumento que converge para seu ponto de vista.

Voltando-se às ocorrências (01) e (02), observamos que lá a ação da hipotática não se realiza (mantém-se na negativa) – *saiu sem contar onde iria (saiu, E NÃO contou); saiu sem falar com ninguém (saiu E NÃO falou)*. Contrariamente, em (03) e (04), a ação que se segue após o *sem* é realizada, na medida em que o enunciador *conta e fala* aquilo que não iria mencionar. É nesse jogo argumentativo que emerge um novo plano de significação, em que modo negativo passa a dar lugar a um novo tipo de adição, ou melhor, a um novo tipo de construção, que temos chamado (cf. Lopes, 2025; Lopes e Oliveira, 2024) de MED de acréscimo.

Nesse tipo de construção, a estrutura *sem* + verbo *dicendi* forma um *chunk*, na medida em que as duas formas se tornam uma nova unidade

de sentido, que passa a atuar como um conector na introdução de novos argumentos convergentes (de onde advém a semântica de acréscimo). Nesse tipo de uso, o relevo informativo recai no argumento que sucede o conector¹⁵.

Tal construção é assim representada: [D1 sem V_{dicendi} (que) D2]. Nessa notação, há maior vinculação entre *sem* e verbo *dicendi*, em comparação à notação da construção anterior, em que esses elementos estão separados por colchetes. Além disso, há, de um lado, menor liberdade na seleção de verbos, dado que a noção de acréscimo só ocorre na combinação de *sem* com alguns verbos *dicendi*, como veremos mais à frente; de outro, conforme afirmamos previamente, há diminuição da composicionalidade semântica, uma vez que os sentidos básicos de *sem*, que envolvem negação e ausência, encontram-se esmaecidos para que seja possível inferir um valor de acréscimo dessa construção.

Sob a hipótese de que o MED de acréscimo seria resultado da construcionalização da oração hipotática adverbial (OHA) modal ou condicional negativa introduzida por *sem* + verbo *dicendi*, Lopes (2025) empreendeu um estudo em busca da origem de [D1 sem V_{dicendi} (que) D2] em *corpora* diacrônicos (Vercial, Tycho Brahe e Colónia). Identificou que os primeiros usos desse MED, nos dados analisados, constam do século XVIII, mais especificamente da sequência *sem contar*, em que *contar* é, na verdade, um verbo do domínio da matemática, de sentido análogo a *enumerar*, forma homonímica ao *dicendi contar*. Vejamos algumas ocorrências:

15 Em nossas pesquisas, temos optado pelo termo *acrécimo* no lugar de *adição*, porque optamos por utilizar uma nomenclatura distinta para considerar a diferença entre as relações mais simétricas e mais assimétricas. Enquanto a conjunção coordenativa aditiva pressupõe, via de regra, maior simetria nas relações, no que diz respeito ao relevo da informação, construções como [D1 sem V_{dicendi} (que) D2] conferem maior relevo informativo, ênfase, ao argumento presente na D2.

(05) Na noite de 16 para 17 as 10 horas morreu de 68 anos a 17 de outubro a Duquesa do Cadaval de uma constipação por haver estado ao ar frio que em três dias a matou não fez testamento dizem que por não fazer dano as legítimas das filhas se as renúncias não valerem e tinha segundo se diz vinculado no morgado a sua terça deixou 39 descendentes *sem contar* os bastardos do Duque Dom Jaime e 12 famílias anojadas enterrou-se na capela de Nossa Senhora da Madre de Deus e pegaram no caixão os poucos fidalgos que foram ao enterro eram o Marquês de Alegrete pai o de Valença e o Visconde moço os Condes de Atalaia (...).

(06) A vantagem numérica é a favor das duas potências meridionais, *sem contar* os meios marítimos, seus e da Inglaterra, que lhe podem dar toda a vantagem nas diversões.

(07) Assim é que é muita verdade; mas também é muita verdade que os Romanos, que não cediam aos Gregos, empreenderam e conseguiram ganhar palmas, em diversos empregos: *sem contar* Virgílio, que três cetros empunhou, na Eneida, Geórgicas, e Éclogas, já Cícero tinha antes de nele sido aclamado por Orador, sumo, óptimo Filósofo, excelente escritos de Diálogos (...).

Nas duas primeiras ocorrências, *sem contar* veicula a noção de acréscimo, advinda, sobretudo, do sentido enumerativo prototípico do verbo *contar*, que coocorre ao lado de outros elementos de natureza também enumerativa: 39 descendentes, 12 famílias anojadas, vantagem numérica etc. Em (07), por sua vez, embora a noção enumerativa ainda possa ser atribuída a *contar*, ela é um pouco mais opaca, uma vez que as informações mais salientes em D1 e D2 não implicam diretamente sobre o verbo uma semântica de *soma* ou *enumeração*, mas, sim, de *relato*.

Lopes (2025) defende que a opacidade presente em ocorrências como (07) permitiu a posterior analogização de novos *types* de MED de acréscimo, identificados nos *corpora* somente a partir do século XX, como *sem falar*, *sem*

dizer, sem mencionar. Nesse caso, teria havido o seguinte tipo de associação: *contar* está para *falar*, assim como *sem contar* está para *sem falar*. Podemos ver duas ocorrências desses novos *types* a seguir:

(08) A parte mais interessante de sua mensagem *sem falar* nas fascinantes evocações históricas que ela contém, é talvez a em que há uma ponderada advertência contra os perigos do nacionalismo.

(09) Desonestidade é sinal de fraqueza. *Sem mencionar que* é a pior estratégia de marketing que existe.

Apesar da expansão de sua produtividade, Lopes (2025) também observou que há verbos de pouca ou nenhuma produtividade na construção de MED. Na análise de 5.320 ocorrências, extraídas do *Corpus Web 2020*, da plataforma *Sketch Engine*, o autor identificou que os verbos *especificar* e *revelar* têm baixíssima produtividade na construção de MED (02 e 01 ocorrência, contra 293 e 322 usos de *sem especificar* e *sem revelar* na construção de OHA) e *explicar* é improdutivo (todas as 270 ocorrências de *sem explicar* são instanciações de OHA).

Feita essa introdução sobre o uso, a produtividade e a gênese do MED de acréscimo [D1 sem V_{dicendi} (que) D2], realizamos, neste trabalho, uma análise colexêmica distintiva com o objetivo de verificar as diferenças sutis entre construções aparentemente sinônimas. Para esse fim, levantamos todas as ocorrências de *sem* + verbo *dicendi* presentes no *Portuguese Corpus 2022*, do Wordschatz Leipzig, constituído exclusivamente de notícias eletrônicas em português. O *corpus* tem um total de 19.461.242 palavras e seus dados foram processados pelo *Software AntConc*.

No *corpus*, identificaram-se 35 lexemas, originalmente *dicendi* ou que podem atuar como verbo de elocução em determinados contextos, mas apenas 22 deles se adjungiam à preposição *sem*. Essas informações constam da Tabela 4.

Lexema	N	Sequência	N
contar	2.213	sem contar	187
falar	3.901	sem falar	154
mencionar	164	sem mencionar	34
dizer	3643	sem dizer	29
citar	328	sem citar	99
comentar	481	sem comentar	5
revelar	380	sem revelar	20
explicar	699	sem explicar	7
especificar	66	sem especificar	46
considerar	637	sem considerar	47
apresentar	2.544	sem apresentar	52
adiantar	102	sem adiantar	11
avisar	140	sem avisar	15
informar	702	sem informar	18
relatar	145	sem relatar	0
perguntar	281	sem perguntar	3
responder	1.229	sem responder	9
indagar	15	sem indagar	0
replicar	42	sem replicar	1
exclamar	3	sem exclamar	0
interrogar	7	sem interrogar	0
negar	265	sem negar	3
afirmar	526	sem afirmar	0
contestar	78	sem contestar	0
argumentar	67	sem argumentar	0
determinar	412	sem determinar	0
solicitar	693	sem solicitar	0
ordenar	40	sem ordenar	0
pedir	1624	sem pedir	6
aconselhar	16	sem aconselhar	0
admitir	153	sem admitir	1
suplicar	02	sem suplicar	0
recusar	136	sem recusar	0
concordar	124	sem concordar	0
questionar	281	sem questionar	6
Total	22.139	Total	754

Tabela 4. Total de verbos *dicendi* no Portuguese Corpus 2022, sozinhos e justapostos a *sem*.
 Fonte: elaboração própria.

Das 754 ocorrências em que o verbo *dicendi* se justapõe a *sem*, a distribuição entre as construções de OHA e de MED se deu conforme indicado na Tabela 5.

SEQUÊNCIA	OHA	MED
sem contar	29	158
sem falar	66	88
sem citar	96	03
sem mencionar	24	10
sem dizer	28	01
sem revelar	20	0
sem explicar	07	0
sem comentar	05	0
sem especificar	46	0
sem considerar	46	1
sem apresentar	52	0
sem adiantar	11	0
sem avisar	15	0
sem informar	18	0
sem perguntar	03	0
sem replicar	01	0
sem responder	09	0
sem negar	03	0
sem solicitar	01	0
sem pedir	06	0
sem admitir	01	0
sem questionar	06	0
Total	493	261

Tabela 5. Distribuição de *sem* + verbo *dicendi* conforme instancicações de OHA e de MED.
 Fonte: elaboração própria.

Os dados expressos na Tabela 5 foram utilizados para a realização da ACD. Para esse fim, importamos a Tabela 5 para dentro do *Programa R* e

rodamos o comando *collex.dist()* dentro do pacote *collostructions*, desenvolvido por Susanne Flach¹⁶. Na próxima seção, discutimos os resultados dessa análise.

3. Resultados e discussão

Abaixo, segue a tabela com os resultados da ACD, obtidos no *Programa R*, dentro do pacote *collostructions*:

COLEX	O.MED	E.MED	O.OHA	E.OHA	ASSOC	F.COL	SIGNIF	COMP
contar	158	64,7	29	122,3	MED	273,95814	*****	SIM
falar	88	53,3	66	100,7	MED	41,59115	*****	SIM
mencionar	10	11,8	24	22,2	OHA	0,43679	ns	SIM
replicar	0	0,3	1	0,7	OHA	0,85047	ns	NÃO
solicitar	0	0,3	1	0,7	OHA	0,85047	ns	NÃO
admitir	0	0,3	1	0,7	OHA	0,85047	ns	NÃO
perguntar	0	1,0	3	2,0	OHA	2,55564	ns	NÃO
negar	0	1,0	3	2,0	OHA	2,55564	ns	NÃO
comentar	0	1,7	5	3,3	OHA	4,26648	*	NÃO
pedir	0	2,1	6	3,9	OHA	5,12405	*	NÃO
questionar	0	2,1	6	3,9	OHA	5,12405	*	NÃO
explicar	0	2,4	7	4,6	OHA	5,98304	*	NÃO
responder	0	3,1	9	5,9	OHA	7,70535	**	NÃO
adiantar	0	3,8	11	7,2	OHA	9,43345	**	NÃO
avisar	0	5,2	15	9,8	OHA	12,90718	***	NÃO
informar	0	6,2	18	11,8	OHA	15,52797	****	NÃO
revelar	0	6,9	20	13,1	OHA	17,28263	****	NÃO
dizer	1	10,0	28	19,0	OHA	17,71062	****	SIM
considerar	1	16,3	46	30,7	OHA	32,97560	*****	SIM
especificar	0	15,9	46	30,1	OHA	40,65619	*****	NÃO
apresentar	0	18	52	34	OHA	46,20473	*****	NÃO
citar	3	34,3	96	64,7	OHA	67,52318	*****	SIM

Tabela 6. Resultados da análise colexêmica distintiva de *sem* + verbo *dicendi*. Fonte: elaboração própria.

16 Disponível em: <https://sfla.ch/collostructions/> - Acesso em 20 mar. 2025.

A Tabela 6 é dividida em 9 colunas, que trazem as seguintes informações: 1) colexemas que se justapõem a preposição *sem*; 2) frequência observada da construção de MED; 3) frequência esperada da construção de MED; 4) frequência observada da construção de OHA; 5) frequência esperada da construção de OHA; 6) associação da construção, se com OHA ou MED; 7) força colostrucional; 8) Nível de significância; 9) se o lexema é compartilhado ou não entre as duas construções.

Uma vez que a ACD busca analisar se existe ou não uma maior associação entre um lexema e uma construção específica, essa medida é calculada com base na relação entre a frequência observada (dados empíricos atestados em *corpus*) e a frequência esperada (quantidade numérica média de ocorrências se os lexemas fossem empregados pelas duas construções por mero “acaso”, isto é, sem uma força de atração específica). Quando essas duas frequências são relativamente próximas, os resultados do nível de associação são considerados não significativos. A título de ilustração, podemos mencionar os colexemas *mencionar*, *replicar*, *solicitar*, *admitir*, *perguntar* e *negar*, que, a despeito de o sistema tê-los associado à construção de OHA, os resultados não foram considerados significativos, haja vista que as frequências observadas e esperadas são relativamente próximas. No caso de *contar* e de *falar*, contrariamente, sua associação com MED é bastante significativa, dado que a frequência observada para *contar* e *falar* na construção de MED (158 e 88, respectivamente) é 244% maior do que a frequência esperada para *contar* (64,7) e 165% maior para *falar* (53,3).

A força colostrucional é uma medida logarítmica, conhecida como *-2log likelihood*, empregada em modelos estatísticos que envolvem probabilidade. Ela busca aferir o grau de associação entre a palavra e a construção. Nesse sentido, quanto maior o valor, mais forte e estatisticamente significativa é a força de associação entre o lexema e a construção. O cálculo é feito com base na frequência bruta do lexema nas duas construções e na frequência geral de todos os lexemas previstos pelo modelo de análise. Podemos verificar, com base nos dados da Tabela 6, que *contar* é muito fortemente atraído para a

construção de MED (F.COL= 273,95). Dada a sua força colostrucional em comparação à outra construção atraída para a construção de MED (*sem falar*), podemos concluir que ele é o lexema prototípico da construção – de maneira análoga ao que observamos, na seção dos pressupostos teóricos, na relação existente entre o verbo *give* e a construção ditransitiva. O lexema mais atraído por OHA é *citar* (F.COL=67,52), mas, uma vez que essa força não é tão maior que outros lexemas presentes no modelo – e.g.: *apresentar* (F.COL=46,24), *especificar* (F.COL= 40,65), *considerar* (F.COL=32,97) –, é provável que não haja, para OHA, um lexema realmente prototípico. Como discutimos previamente, a priori, qualquer verbo (nocional ou não nocional) pode se adjungir a *sem* na construção de OHA, e não somente verbos *dicendi*.

Os níveis de significância buscam verificar se a associação estabelecida pelo modelo pode ser considerada estatisticamente significativa, cujos resultados são calculados com base no Teste Exato de Fisher. NS quer dizer *não significativo*, o que indica que as chances de erro no resultado de associação são superiores a 5% ($p > 0,05$). Os asteriscos (ou estrelas), por sua vez, já consideram que o resultado é estatisticamente significativo e devem ser lidos da seguinte maneira: *, $p < 0,05$; **, $p < 0,01$; ***, $p < 0,001$; ****, $p < 0,0001$; *****, $p < 0,00001$). Sob esse ponto de vista, atingem nível de significância os resultados atestados para *contar*, *falar*, *considerar*, *especificar*, *apresentar* e *citar*.

Por fim, os lexemas podem ser compartilhados entre as duas construções, mesmo quando o nível de associação é mais forte para uma delas. É o que observamos para *contar*, *falar*, *mencionar*, *dizer*, *considerar* e *citar*. Para que o compartilhamento seja considerado positivo, basta que haja pelo menos uma ocorrência atestada em cada uma das construções. Portanto, quando o resultado do compartilhamento for “não”, há uma forte tendência de repulsão entre o lexema e a construção para a qual o modelo não atestou associação.

Relacionando-se os resultados obtidos pela ACD com as investigações prévias sobre a emergência das construções de MED de acréscimo formadas por *sem* + verbo *dicendi*, podemos chegar a algumas conclusões, quais sejam:

a) o verbo contar, seja a forma de sentido enumerativo, seja a *dicendi*, continua sendo o lexema prototípico na construção de MED.

Nesse caso, sua força colostrucional muito discrepante em relação a todos os outros resultados obtidos sugere que já há uma grande associação entre essa sequência de elementos – *sem* + *contar* – e a construção de MED, de modo que, quando um falante mobiliza essas estruturas para o uso, normalmente o faz para estabelecer relações de acréscimo. Dentro de um modelo de descrição linguística com forte aparato cognitivo, como o nosso, entendemos que, à medida que as sequências linguísticas vão se rotinizando, por meio de processos cognitivos de automatização, em determinados contextos de uso, essas estruturas se fixam mais na memória e, conseqüentemente, se tornam mais facilmente acessíveis.

Sob esse ponto de vista, a maior acessibilidade de *sem contar* para a construção de MED tende a tornar mais custoso o seu processamento cognitivo nas construções de OHA. Isso sugere que pode haver uma tendência de, para expressar sentidos modais ou condicionais negativos, o falante recorra a outros verbos, mesmo sendo *contar* um verbo de significação bastante geral.

Corroborando com essa hipótese, inclusive, os resultados obtidos para *sem considerar*, *sem especificar*, *sem apresentar* e *sem citar*, em comparação com *sem contar*, haja vista que eles apresentam uma frequência observada na construção de OHA bastante superior (46, 46, 52 e 96, respectivamente) a *sem contar* (29), mesmo que *sem contar* seja a sequência mais frequente em todas as construções analisadas no *corpus* (187 ocorrências).

b) a construção vem expandindo seus types, progressivamente, priorizando, aparentemente, verbos de significação mais geral:

Depois da emergência de *sem contar* (atestado em *corpora* no século XVIII, como vimos previamente), novos usos teriam surgido no português por meio de analogização, provavelmente no século XX (conforme também atestado em *corpora*), sendo o primeiro deles *sem falar*, que é o verbo *dicendi* de significação geral mais presente no *corpus* (3.901 ocorrências de *falar*, no infinitivo, em todo o *corpus*). Os demais lexemas, no entanto, embora possam ocorrer na construção de MED, como *mencionar*, *dizer*, *considerar* e *citar*, ainda são mais fortemente atraídos para OHA.

Esse fenômeno de expansão de *types*, conhecido como expansão *host-class* (Himmelmann, 2004), é característico de mudanças pós-construcionalização (cf. Traugott; Trousdale, 2013). Isso significa que, à medida que MEDs de acréscimo formados por *sem contar*, *sem falar* etc. vão aumentando sua frequência de uso, mais os falantes promovem novas analogizações com outros lexemas, desde que eles atendam as propriedades semânticas daqueles que já normalmente preenchem os *slots* da construção.

O que nos causa um certo estranhamento nos dados é entender por que *dizer* é menos atraído para a construção de MED, uma vez que ele é um verbo de significação geral altamente frequente. No *corpus* investigado, por exemplo, é o segundo verbo *dicendi* mais frequente depois de *falar*: 3.643 ocorrências. A nossa hipótese é que essa menor produtividade possa estar associada a uma maior vinculação de *sem dizer* às construções de OHA, o que vai ao encontro do argumento de que isso seja uma consequência do bloqueio estatístico (cf. Goldberg, 2019). Nesse caso, valem as mesmas reflexões que utilizamos para *sem contar*: se o falante já mobiliza essa estrutura recorrentemente para a expressão de modo ou de condição, essas propriedades ficam registradas mais fortemente na memória e, conseqüentemente, mais acessíveis.

c) Verbos *dicendi* de significação mais específica tendem a restringir-se às construções de OHA

Verbos que indicam um tipo distinto, mais específico de elocução, como *revelar*, *especificar*, *explicar*, *perguntar*, *negar*, *pedir*, *questionar*, *responder* etc., quando adjungidos à preposição *sem*, estão normalmente subordinados a um outro verbo *dicendi*, isto é, já especificam uma cena enunciativa presente na oração matriz, modalizando seu verbo. Por isso, tais lexemas estão fortemente associados às construções de OHA. Como ilustração, vejamos três ocorrências:

(10) No grupo, elas emitiam opiniões predominantemente sobre os temas, *sem revelar* fatos de suas próprias vidas.

(11) Greve dos correios não prejudicará Enem, diz Haddad. Ministro defendeu taxa de participação mínima das escolas, *sem especificar* percentual.

(12) Enunciando estas categorias *sem explicar*, por razões de brevidade, Hugo retoma Isidoro.

Nas ocorrências (10), (11) e (12), as construções de OHA têm como escopo os verbos e expressões de elocução presentes na D1, a saber: *emitiam opiniões*, *defendeu* e *enunciando*, respectivamente. Lopes (2025) já havia identificado que, enquanto verbos de significação mais geral, como *contar*, *falar*, *dizer*, nos usos de OHA, escopam verbos de significação variada em D1 (para além de outras formas *dicendi*) – como observamos nas ocorrências (01) e (02), por exemplo, *saiu sem contar onde ia*; *saiu sem falar com a imprensa* –, verbos de significação específica, nas construções de OHA, circunstanciam o significado de um verbo ou expressão de elocução em D1: *disse sem revelar*; *contestou sem perguntar*, *contou sem responder* etc. Sob esse ponto de vista, esses verbos mostram-se improdutivos para a construção de MED no *corpus* analisado. Logo, nessa esteira, embora os verbos *dicendi* como um todo respondam ao princípio da coerência semântica, conforme advoga

Goldberg (1995), a construção de MED ainda prioriza a seleção de verbos de significação mais geral.

Considerações finais

Neste texto, propusemos que a análise colexêmica distintiva (ACD), originalmente empregada em perspectiva sincrônica, pode ser uma boa ferramenta de análise para o estudo da construcionalização gramatical em construções complexas. Apesar de essa não ser uma proposta inovadora, na medida em que Hilpert (2006) já havia evidenciado que a ACD pode trazer contribuições para mapear diferenças de produtividade das construções ao longo do tempo, reconhecemos que a aplicação que fizemos neste artigo é distinta, na medida em que olhamos para duas construções complexas – A e B –, em que a segunda é o resultado da recategorização da primeira.

Como já têm atestado diversos estudos em mudança, as construções procedurais resultantes de construcionalização gramatical apresentam alguns aspectos afins, que envolvem: a) o aumento de vinculação entre as subpartes da construção – em decorrência de processos de automatização e *chunking*; b) diminuição de sua composicionalidade semântica e/ou sintática; c) restrições colocacionais¹⁷. Sob esse ponto de vista, a comparação entre os lexemas associados às construções A e B, dentro do modelo da ACD, pode nos ajudar a entender e descrever um modelo de conhecimento linguístico estruturado em rede, em que o aumento de produtividade *type* e *token* de uma construção B pode impactar diretamente as mesmas produtividades em A.

A esse respeito, no que tange à construção de MED [D1 sem V_{dicendi} (que) D2], os dados sugerem que a elevada frequência do lexema *contar* dentro dessa construção impacta diretamente a sua instanciação na construção de

17 Temos observado que os parâmetros de gramaticalização de Lehmann (1995), pensados para a perspectiva da gramaticalização, funcionam muito bem para a análise e descrição da construção gramatical complexa, cuja emergência tende a sofrer os mesmos processos: atribuição, paradigmatisação, obrigatoriedade, condensação, coalescência e fixação.

OHA. Isso decorre do fato de as sequências linguísticas mais automatizadas, recorrentemente recrutadas para o uso em determinados contextos, fixarem-se mais fortemente na memória e, conseqüentemente, tornarem-se mais acessíveis. Muito provavelmente, daí decorre o fato de a frequência observada de *sem contar* para OHA ser muito menor do que a esperada, como discutimos anteriormente, durante a análise da Tabela 6.

Desse modo, uma vez que as línguas são dinâmicas e estão continuamente sujeitas a processos de instabilidade, a tendência é que essas relações sequenciais – previstas nas construções A e B – sejam continuamente reconfiguradas. Sendo assim, é esperado que, com o passar do tempo, a construção B produza novos *types*, que, atingindo frequência suficiente, devem impactar tanto a representação das construções de B, em termos de sua própria produtividade *type* e *token*, quanto as de A.

Referências

LOPES, M. G.; MOURA, S. C. [Sem Vdicendi QUE]: um conector hipotático de adição do português. *PERcursos Linguísticos*, v. 12, n. 30, p. 235-255, 2022.

LOPES, M. G.; OLIVEIRA, B. P. De oração modal ou condicional a operador argumentativo de acréscimo: o papel da intersubjetividade na emergência de [sem falar] no português. *Revista Linguística/Revista do Programa de Pós-Graduação em Linguística da Universidade Federal do Rio de Janeiro*, v. 20, n.1, p. 48-70, 2024.

LOPES, M. G. Marcadores Estruturantes do Discurso instanciados pelo subesquema [D1 (isso) sem Vdicendi (que) D2] no português brasileiro contemporâneo: um estudo centrado no uso. *Relatório de Estágio de Pós-Doutoramento*. 112f. Faculdade de Letras, Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2025.

BYBEE, J. L. Mechanisms of change in grammaticalization: the role of Frequency. In: JOSEPH, B. D.; JANDA, J. (eds.). *The handbook of Historical Linguistics*. Oxford, Blackwell, 2003, p. 602-623.

BYBEE, J. L. *Language, Usage and Cognition*. New York: Cambridge University Press, 2010.

CROFT, W. *Radical Construction Grammar*. New York: Oxford University Press, 2001.

CUNHA, C. E.; CINTRA, L. F.L. *Nova Gramática do Português Contemporâneo*. 7ª. Ed. Rio de Janeiro: Lexikon, 2001.

DIESEL, H. *The Grammar Network*. How linguistic structure is shaped by language use. New York: Cambridge University Press, 2019.

GOLDBERG, A. *A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press, 1995.

HILPERT, M. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2, n. 2, p. 243-256, 2006.

HILPERT, M. *Construction Grammar and its application to English*. Edinburgh: Edinburgh University Press, 2014.

HIMMELMANN, N. Lexicalization and grammaticization: opposite or orthogonal? In: BISANG, W.; HIMMELMANN, N.; WIEMER, B. *What makes grammaticalization – A Look from its Fringes and its Components*. Berlin: Mouton de Gruyter, 2004, p. 21-42.

LANGACKER, R. *Foundations of Cognitive Grammar*. Vol 1. Theoretical Prerequisites. Stanford: Stanford University Press, 1987.

NEVES, M. H. M. *A gramática do português revelada em textos*. São Paulo: UNESP, 2018.

RAPOSO, E. B. P. *et al.* (orgs.). *Gramática da Língua Portuguesa*. Vol I. Lisboa: Fundação Calouste Gulbenkian, 2015.

ROCHA LIMA, C. H. *Gramática Normativa da Língua Portuguesa*. 1ª. Ed. Rio de Janeiro: José Olympio, 1972.

STEFANOWITCH, A.; FLACH, S. “Too big to fail but big enough to pay for the mistakes”: A collostructional analysis of the patterns [too ADJ to V] and [ADJ enough toV]. In: Gloria Corpas & Jean Pierre Colson (eds.). *Computational and corpus-based phraseology*. Amsterdam: John Benjamins, p. 248-272, 2020.

STEFANOWITSCH, A.; GRIES, S. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, v. 8, n.2., p. 209-243, 2003.

STEFANOWITCH, A.; GRIES, S. Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, v. 1, n.1, p.1-43, 2005.

STEFANOWITCH, A.; GRIES, S. Extending collostructional analysis. A corpus based perspective on alternations. *International Journal of Corpus Linguistics*, v. 9, n.1, p. 07-129, 2004.

TRAUGOTT, E. C. *Discourse Structuring Markers in English*. Philadelphia: John Benjamins, 2022.

TRAUGOTT, E. C. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 57/1, p. 33-65, 1989.

TRAUGOTT, E. C.; DASHER, R. *Regularity in Semantic Change*. Cambridge: Cambridge University Press, 2002.

TRAUGOTT, E. C.; TROUSDALE, G. *Constructionalization and constructional changes*. Oxford: Oxford University Press, 2013.