DOI: 10.14393/DL52-v16n4a2022-2

A Gramateca e a Literateca como macroscópios linguísticos

Gramateca and *Literateca* as language macroscopes

Diana SANTOS[*]

RESUMO: Neste artigo exploramos várias potencialidades que os ambientes da Gramateca e da Literateca permitem aos usuários interessados na pesquisa em língua portuguesa. Por um lado, apresentamos estes ambientes dando conta de novas funcionalidades acessíveis; por outro, trazemos dez exemplos de perguntas de pesquisa para demonstrar a utilidade da existência destes serviços, que pretendem ser uma espécie de macroscópio para observar a língua, nas vertentes semântica e morfossintática, assim como para a leitura distante de textos literários e a extração de informação em português.

PALAVRAS-CHAVE: Corpos. Visualização. Linguística com corpos. Estudos literários. Português.

ABSTRACT: This paper demonstrates several features of *Gramateca* and *Literateca*, which are environments for linguistic and literary research on top of being a large richly annotated corpora in Portuguese. The paper presents them and pinpoints several new functionalities; it additionally provides ten examples of research questions that demonstrate the usefulness of this kind of Web-based service, that can be conceived as language macroscopes. Examples concern semantics, morphosyntax, distant reading of literary texts, and information extraction.

KEYWORDS: Corpora. Visualization. Corpus linguistics. Literary studies. Portuguese.

1 Introdução

Há cerca de 20 anos a Linguateca proporciona, a todos os linguistas e cientistas da computação interessados no processamento do português, uma plataforma para buscar e obter dados linguísticos, através do projeto AC/DC, desde a publicação dos artigos seminais Santos e Ranchhod (1999) e Santos e Bick (2000). Ao longo do tempo,

[*] Doutora, Universidade de Oslo e Linguateca. ORCID: https://orcid.org/0000-0002-3108-7706. d.s.m.santos@ilos.uio.no

muito mais material de vários gêneros foi incorporado, graças à adesão da comunidade, o que nos permitiu disponibilizar e tratar corpos[1] não criados por nós, sempre em regime de não exclusividade. Da mesma forma, vários grupos uniram esforços com a Linguateca para criar novos recursos em parceria, muitos dos quais em constante crescimento, como é, por exemplo, o caso do OBras (SANTOS *et al*., 2018). Isso implica que as descrições do material e dos recursos acessíveis necessitem de alguma atualização desde o artigo Santos (2014), que buscou oferecer um histórico da atividade com corpos na Linguateca. Além disso, devido ao nosso constante uso dos corpos, é de se esperar que tenhamos desenvolvido também novas informações a eles associada, e novas formas de as visualizar, que merecem ser descritas à comunidade linguística e de processamento de linguagem natural (PLN).

2 Gramateca

A Gramateca, conceitualizada em 2014, foi apresentada em Santos (2014) como um ambiente para estudar a gramática da língua portuguesa. Ela contém, além dos corpos propriamente ditos, a potencialidade de criar formulários de interação com vários falantes para testar casos complicados, através do Rêve – unindo, assim estudos quantitativos com qualitativos. Santos *et al*. (2015) descreve em mais detalhe essa nova forma de fazer gramática, que utiliza textos reais, diferentes gêneros, e, adicionalmente, agrega a capacidade de levar em conta diversos informantes. Esse artigo exemplificou os conectores condicionais, a descrição fina do corpo humano, e os diferentes sentidos do verbo *admirar.*

Mas outras questões foram também sendo identificadas como pertinentes para uma visualização do conteúdo de um corpo, como o artigo de Santos (2015) e a

[1] Para a motivação de usar o termo *corpo* e *corpos* em vez de *corpus* e *corpora*, veja-se Santos (2008).

apresentação de Santos (2018) ilustram: linhas de tempo, formas distintas de identificação de uma mesma entidade, nuvens de palavras etc.

O que significa que, associadas à Gramateca, se encontram várias ferramentas que permitem manipular e apresentar resultados de forma mais intuitiva. Tais ferramentas ajudam o pesquisador a explorar o material e serão exemplificadas na sequência, depois de apresentarmos a Literateca.

3 A Literateca

De fato, foi o início da consideração dos textos literários como um gênero por si só, com requisitos específicos, e com bastante informação associada a cada texto (os metadados, que incluem autor, data, gênero do autor, gênero literário, escola literária, canonicidade...), que levou a uma explosão de novas ferramentas associadas aos corpos, e que mereceu, portanto o nome de Literateca – que é uma especialização da Gramateca para textos literários.

Assim, além de tudo o que já era possível fazer nos outros corpos, definiram-se distribuições por período, assim como a anotação de localizações, e de personagens, permitindo a construção de redes de personagens e da sua presença na obra ao longo do tempo, como foi descrito em Santos e Freitas (2019), e o desenho automático de mapas, cf. Santos e Alves (2022). Foi também possível utilizar ferramentas de redução de dimensionalidade para oferecer uma visão de conjunto usando muitas características diferentes, como ilustrado em Santos *et al*. (2020).

É importante, contudo, deixar claro que os desenvolvimentos na Literateca são consequência do que já existe na Gramateca, e que algumas anotações iniciadas na Literateca foram depois expandidas a outros corpos, ou seja, ultrapassaram o domínio restrito dos textos literários. E, de qualquer maneira, a existência da Literateca como um subconjunto da Gramateca com características especiais se beneficia grandemente da comparação com os corpos todos, para identificar o que é sobretudo literário ou

pertence à língua em geral. Esperamos poder ilustrar isso de forma convincente no resto do artigo.

4 Algumas explorações em textos literários

Começaremos por perguntas relativamente fáceis, e vamos aumentando a sofisticação linguística, de forma a vermos como estes ambientes, ou infraestruturas de pesquisa, permitem estudar o português e o português brasileiro em especial.

Mas, não podemos deixar de destacar que, mesmo que as perguntas sejam fáceis, está em geral subentendido muito trabalho, além da reflexão sobre a anotação à qual as perguntas recorrem.

Por outro lado, é sempre necessário não se esquecer a base sobre a qual se pergunta e que determina o universo explorado. Em relação aos textos literários, e devido à problemática dos direitos autorais, contamos na sua maioria apenas com obras já no domínio público, e daí as perguntas literárias que faremos se referirem a livros publicados há mais de 80 anos. Apresentamos, a seguir, as perguntas de pesquisa.

#1 *Quais autores do século XIX incluem mais nomes de localizações nos seus textos?*

O que é um autor do século XIX? Escolhemos aqui operacionalizar esta questão como "autores de obras publicadas no século XIX", e só contar com essas obras, mas evidentemente outras escolhas seriam possíveis. Veja-se, a seguir, a Tabela 1, na qual são listados autores do século XIX e a incidência de topônimos em suas obras.

Tabela 1 – A presença de topónimos em autores com obras publicadas no século XIX.

Autor	**Loc.**	**Palavras**	**loc/pal*1000**
Marquês de Fronteira e d'Alorna	1236	67 031	18,44
Manuel Pinheiro Chagas	1103	77 178	14,29
Zeferino Norberto Gonçalves Brandão	862	60 869	14,16
Joaquim Nabuco	1073	80 144	13,39
Alexandre Herculano	17867	1 484 356	12,04
Ramalho Ortigão	788	67 535	11,67
Alexandre de Serpa Pinto	2167	192 543	11,25
Alberto Pimenta	1743	158 688	10,98
Teixeira de Vasconcellos	1033	95 742	10,79
António Nobre	295	28 936	10,19
Matilde Isabel de Santana e V. M. Bettencourt	423	42 502	9,95
Adolfo Caminha	1911	193 735	9,86
Luciano Cordeiro	1498	160 322	9,34
António Francisco Barata	1850	198 275	9,33
Candido de Figueiredo	166	17 871	9,29
Bernardino Pereira Pinheiro	465	50 803	9,15
Alfredo Campos	291	33 916	8,58
Tomaz de Melo	539	64 151	8,4
Inglês de Sousa	996	119 746	8,32
Inácio Pizarro de Morais Sarmento	346	45 312	7,64

Fonte: elaborada pela autora.

Uma rápida análise dos resultados na Tabela 1 permite-nos observar que os romances históricos, e aqueles que se passam em zonas em princípio desconhecidas do leitor, como é o caso da Amazônia para Inglês de Sousa, são os que têm maior densidade toponímica. A presença de muito mais autores portugueses do que brasileiros nesta lista deve-se ao fato de a própria amostra apenas conter (por enquanto) 28 autores brasileiros deste período, num universo de 117 autores no total.

#2 Quais obras brasileiras têm mais referências a etnicidade?

Se procurarmos agora apenas entre as 203 obras brasileiras presentes na Literateca, a Tabela 2 ilustra as que têm mais menções ao campo semântico da etnicidade.

Tabela 2 – A presença da etnicidade em obras brasileiras.

Obra	Etn.	Pal.	Freq. Rel. em 10000 palavras
Rei negro	430	67053	64,13
Uma tragédia no Amazonas	114	28530	39,95
O Uraguai	57	14684	38,81
O Guarani	449	124417	36,08
As Vítimas-Algozes	374	112716	33,18
Nove noites	16	5077	31,51
Banzo	99	32136	30,81
A escrava	16	5274	30,34
Iracema, lenda do Ceará	102	34788	29,32
O sonho das esmeraldas	81	30375	26,67
Pausa	2	754	26,53
Os irmãos Leme	91	35276	25,8
A viagem maravilhosa	335	133116	25,17
Vidros quebrados	5	2004	24,95
Turbilhão	217	88164	24,61
O Bom-Crioulo	110	45303	24,28

Fonte: elaborada pela autora.

Se em alguns casos das 16 obras com proporcionalmente mais termos relativos a etnicidade isso seria previsível pelo seu título, convém chamar a atenção para os dois contos com apenas duas ou cinco palavras remetendo para esse domínio, de Moacyr Scliar e de Machado de Assis respetivamente, que chegaram à lista devido ao seu tamanho reduzido. Para evitar isto, poderíamos ter escolhido apenas romances e novelas, mas este é mais um caso em que a pergunta poderia ter sido interpretada de maneira diferente.

#3 *Entre médicos e boticários por um lado, e padres e frades por outro, quais são as profissões mais presentes na literatura lusófona?*

Mais uma vez, esta pergunta pode ser operacionalizada de diferentes formas: simplesmente contando as vezes que cada uma das profissões aparece no total das obras, ou identificar esse assunto por obra: quantas obras só têm referência a uma profissão ou a outra, e em quantas obras uma supera a outra?

Num universo de 860 obras, em 472 obras há referência a um médico ou boticário, enquanto em 582 há referência a padres, frades, freiras ou monges.

E a conclusão é inescapável. Embora a medicina e a profissão médica sejam muito presentes na literatura lusófona – veja-se Santos (2019) e Langfeldt (2021) –, na amostra que temos a profissão religiosa ainda é mais conspícua. Com efeito, das 587 obras em que uma ou ambas as ocupações, ou vocações, ocorrem, 96% (561 obras) falam mais da classe religiosa, sejam heróis ou vilões. É também interessante notar que apenas 92 obras mencionam ambos os tipos de profissões.

Um exemplo que demonstra inequivocamente este predomínio é a obra *A Mortalha de Alzira.* Sendo esta a obra que contém mais referências às palavras *médico* e *boticário* em valores absolutos, 40, apresenta mesmo assim 54 menções a palavras descrevendo religiosos.

#4 Quais são as ações mais associadas a uma ou a outra profissão?

Em 1.133 casos de um médico sujeito (de uma frase ativa ou passiva)[2], os seguintes verbos são mais frequentes do que para a média dos sujeitos: *receitar*, *recomendar*, *aconselhar*, *declarar, examinar*, *chamar* e *tomar*. Em 3.020 casos de um religioso sujeito (de uma frase ativa ou passiva), são os seguintes os verbos mais frequentes do que a média: *curar*, *rezar*, *erguer-se, levantar* e *olhar*.

Enquanto alguns destes verbos concordam imediatamente com a nossa intuição, outros há que merecem investigação: em particular, não conseguimos achar explicação para a alta frequência de *olhar* associada a religiosos, sobretudo se considerarmos que uma quantidade significativa dos casos de *levantar* se refere precisamente a *olhos* ou *olhar*. Já quanto a *tomar*, são as expressões *tomar o pulso*, *tomar notas* e *tomar injeção* que explicam a co-ocorrência preferencial com a profissão médica.

[2] A expressão de busca para obter os verbos foi: [lema="médico|facultativo|boticário|farmacêutico" & pos="N.*"] [pos=".*AUX.*|ADV"]* @[pos="V.*" & pos!=".*AUX.*"]

Mas o mais surpreendente foi o verbo *curar* estar mais associado a religiosos do que a médicos. Ao analisar esses casos, descobrimos um erro da análise sintática, que considerava incorretamente *cura* em *padre cura* como forma do verbo *curar*. Se o mencionamos aqui, é porque é sumamente importante não confiar cegamente nos resultados, e investigar tudo o que vá contra a nossa intuição de falantes de uma língua. Afinal de contas, não é possível garantir que a informação associada a mais de um bilhão de palavras esteja 100% correta. O que fazemos é tentar melhorar, e corrigir semiautomaticamente, os casos que vamos encontrando, usando a filosofia de interação entre pessoas e máquinas apresentada em Santos e Mota (2010).

#5 Em termos de co-ocorrência de emoções, existe distinção entre texto literário e texto jornalístico (por exemplo)?

Dado que a menção de emoções é algo muito corrente em português, talvez a maneira mais fácil de responder a esta pergunta seja tentar visualizar as emoções em conjunto através da sua co-ocorrência, na Literateca, e nos corpos todos, e comparar as duas visualizações.

As figuras (feitas com o programa igraph (CSARDI; NEPUSZ, 2006) do R (R Development Core Team, 2008)) ilustram um grafo de coocorrências entre todas as palavras que estavam marcadas com o campo semântico da emoção, segundo os grupos previamente identificados (veja-se SANTOS; SIMÕES; MOTA (2021) para a documentação deste esforço).

Figura 1 – Grafo das emoções na Literateca.

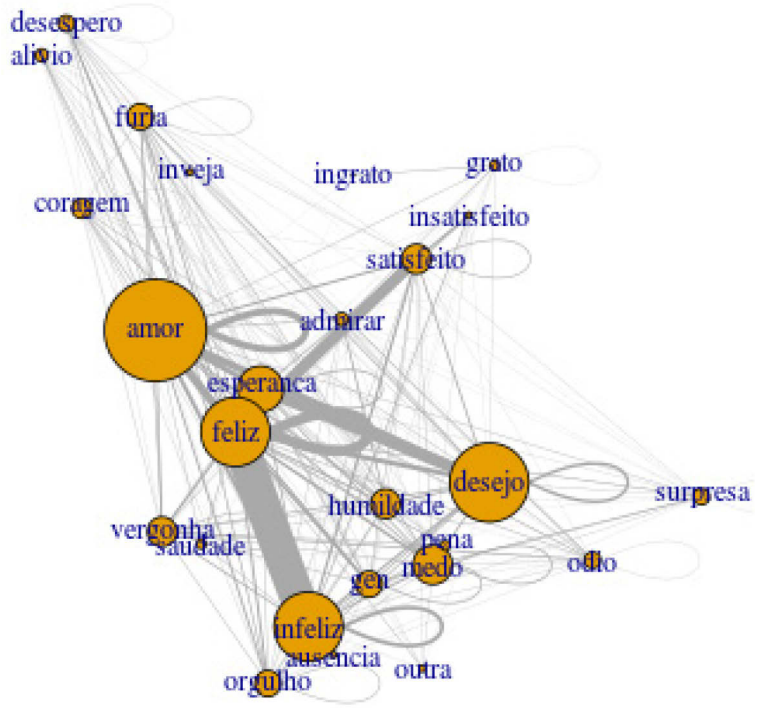
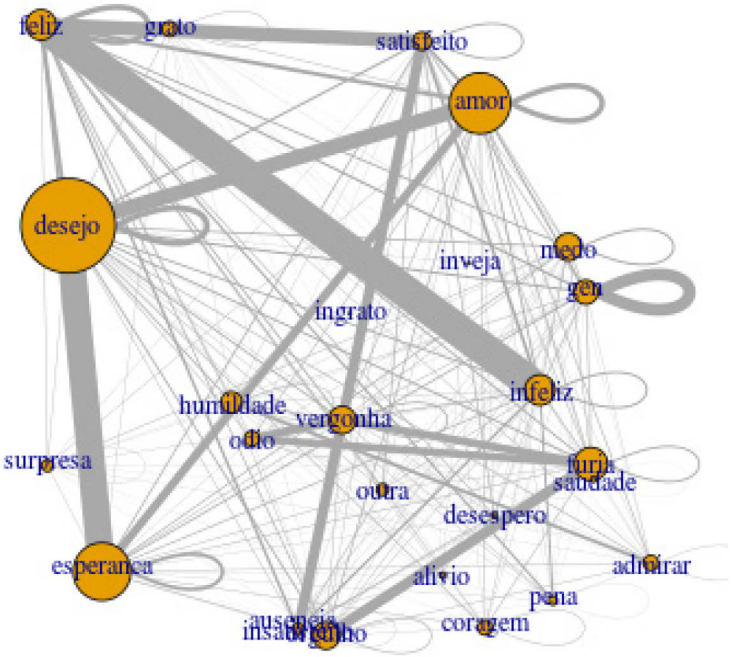


Figura 2 — Grafo das emoções em todos os corpos.



As diferenças entre os dois grafos – que, sendo desenhados aleatoriamente, não mantêm, infelizmente, a mesma posição espacial das emoções – mostram que os

campos semânticos do AMOR e da (IN)FELICIDADE são bastante mais proeminentes nos textos literários (Figura 1), enquanto que nos corpos todos o DESEJO (que aqui seria mais natural chamar intenção e/ou volição) e a ESPERANÇA são mais usados (Figura 2).

Para ilustrar também a riqueza lexical de um destes campos, mostramos a nuvem de palavras associada a AMOR na Figura 3[3].

Figura 3 — Lemas que foram (em contexto) associados à emoção AMOR.

5 Algumas explorações na "língua toda"

Por língua toda queremos indicar a variedade dos corpos todos que temos, e que não incluem, nem podem incluir, evidentemente, toda a língua, mas que contêm

[3] Agradeço a um revisor anônimo pelo seguinte comentário, com que concordo absolutamente: as nuvens de palavras, embora cada vez mais usadas devido ao seu apelo visual, tornam difícil recuperar impressões precisas a partir delas, donde devem ser feitas as devidas ressalvas.

muitos gêneros e tipos de discurso, desde o oral ao religioso, ao acadêmico e à literatura farmacêutica. Feita esta explicação, vejamos então a próxima pergunta:

#6 Que verbos requerem completivas com subjuntivo? Como explicar a variação nos casos em que tanto indicativo quanto subjuntivo são usados?

A questão do uso do subjuntivo, muito frequente em português em comparação com outras línguas românicas, exige que este seja ensinado logo em níveis básicos de língua, e compreendido por quem o ensina. E o que é certo é que existem muitas variáveis que influenciam o seu uso. Em seguida veremos como identificar propriedades relevantes para a escolha de um modo, e criar materiais de ensino em relação às completivas finitas regidas por um verbo (um processo semelhante seria seguido se estivéssemos interessados em completivas regidas por um substantivo, ou por um adjetivo).

Vejamos, por exemplo em texto jornalístico (no corpo CHAVE), qual a distribuição do modo nas completivas finitas associadas aos verbos *esperar*, *acreditar*, *crer*, *considerar*, *julgar*, *achar* e *pensar*[4], indicada na Tabela 3.

Tabela 3 — A distribuição no modo das completivas finitas.

verbo	**SUBJ**	**%**	**IND**	**%**	**Coef. SUBJ/IND**
pensar	149	7.80%	1742	92.12%	0.09
esperar	899	98.60%	12	1.32%	74.9

[4] Um exemplo de como obter estes valores para *achar* seria: [lema="achar"] [word="que"] [pos!="V.*"] [pos="V.*" & temcagr=".*IND.*"] e [lema="achar"] [word="que"] [pos!="V.*"] [pos="V.*" & temcagr=".*SUBJ.*"]

acreditar	380	30.72%	857	69.28%	**0.44**
crer	108	22.60%	370	77.40%	**0.29**
considerar	32	2.46%	1269	97.54%	0.03
julgar	12	4.96%	230	95.04%	0.05
achar	130	3.94%	3165	96.05%	0.04

Fonte: elaborada pela autora.

Vemos que há dois verbos em que existe de fato variação no modo das completivas, nomeadamente *acreditar* e *crer*. Os outros casos ou exigem subjuntivo, como *esperar* (o que era esperado), ou indicativo, embora sejam versões um pouco menos assertivas de *ter certeza (de) que*. A maioria dos casos em que aparece o subjuntivo com esses verbos corresponde ao verbo negado.

Vejamos então com mais atenção os casos de *acreditar* e *crer*, que, embora apresentem o verbo da completiva majoritariamente no indicativo, têm uma fatia considerável no subjuntivo. Por meio da análise dos 488 casos de subjuntivo, verificamos que 271 são negados, 12 correspondem a perguntas, 17 à existência de uma outra oração na completiva (geralmente com *se*) e, portanto, não são casos de subjuntivo devido a *acreditar* ou *crer*, e 30 correspondem a casos da forma *é X acreditar/crer que*, fórmula essa que costuma induzir o subjuntivo (cf. *é bom/provável/difícil que ele venha*) e que pode, portanto, provocá-lo depois do verbo no infinitivo. Outros casos identificados são as expressões *levar a crer*, *ser de crer*, *custar a crer*, assim como casos com o sujeito formado com o pronome *poucos.*

A Figura 4 ilustra a variedade de tempos e modos que são possíveis com a expressão *levar a crer,* nomeadamente o indicativo, o subjuntivo, mas também o futuro e o condicional, o que significa que não faz sentido tentar criar regras para ensinar, visto que as diferenças parecem puramente estilísticas.

Figura 4 — Excerto de concordâncias relativas a *levar a crer que.*

par=FSP951103-063-725: Folha -- Mas tudo **leva a crer que** você é culpada .
par=FSP951116-079-799: Segundo Vieira, tudo **leva a crer que** o ferimento tenha sido feito «com técnica» .
par=FSP950207-008-81: Em face da crise instalada na assistência pública, e sem saber exatamente como ela se processa, o contribuinte pode ser **levado a crer que** terá, com o PAS, algo mais que a desassistência atual .
par=FSP940921-080-811: Na ocasião, Zé Roberto surpreendeu ao anunciar o corte de Pinha, pois tudo **levava a crer que** ele faria a opção entre Jorge Édson e Douglas .

Por outro lado, esta análise mais fina permitiu-nos identificar construções interessantes para serem ensinadas explicitamente, e para treino, por exemplo com o Ensinador (SIMÕES; SANTOS, 2011), um programa de criação de exercícios didáticos baseados nos corpos do AC/DC. Na Figura 5 apresentamos uma possibilidade de exercício, focando o sujeito *poucos* e a construção *é X acreditar*, em que a intenção é que aluno aprenda os verbos apresentados entre parênteses no subjuntivo, preenchendo as lacunas.

Figura 5 — Um exercício associado a completivas objeto de *crer* e *acreditar*.

Coloque o verbo no tempo e modo correto

1. *par=FSP940807-181-2343*: Em ` True Romance ' é impossível acreditar que _____ ele. (ser)
2. *par=PUBLICO-19950311-180-1946*: Era difícil acreditar que _____ passado apenas dez anos desde que Gorbatchov subiu ao poder após a morte de três dinossauros decrépitos. (ter)
3. *par=FSP941106-079-1036*: Poucos acreditam que _____ realista, até porque a inflação será pressionada pela desvalorização do rublo, que encarece as importações. (ser)
4. *par=FSP951219-141-1375*: Penn pode ter amadurecido, mas é difícil acreditar que _____ se livrado das confusões. (ter)
5. *par=PUBLICO-19951128-044-420*: Se é admissível que haja quem 'teja disposto a comprar enchidos ou bebidas a preços baixos, já é difícil acreditar que _____ receptadores de fatias de bolo de chocolate. (existir)
6. *par=FSP950208-008-56*: Com o PIB crescendo 5 % ao ano e o juro real da dívida pública crescendo 35 %, é difícil acreditar que _____ no caminho da «sustentabilidade»! (estar)
7. *par=PUBLICO-19950112-048-328*: Poucos acreditam que _____ sido um negócio destinado a favorecer o «núcleo duro» de accionistas da instituição, apesar de assumirem que o volume e o preço levantam suspeitas. (ter)

#7 Qual a posição preferida de orações participiais?

Outro tópico interessante na sintaxe da língua portuguesa é a existência frequente de orações participiais, ou seja, com o primeiro verbo na forma de particípio passado, e que geralmente caracterizam um elemento da oração, comportando-se como orações adjetivas, mas podendo encontrar-se a distância significativa dele, antes ou depois, como ilustram os exemplos da Figura 6.

Figura 6 — Orações participiais no corpo NILC/São Carlos, mostrando a que entidade se referem.

par=8477: A festa da première de Missão impossível, um dos mais badalados filmes da temporada de verão deste ano — que estréia dia 12 de julho no Brasil — lotou, na noite de segunda-feira, o teatro Mann Village**, comandada** pelo astro principal, Tom Cruise.
par=Mais—94b-1: Mas é preciso ainda lembrar a sua posição a respeito da função social da ciência**, baseada** num ponto de vista democrático que elaborou a partir da sociologia durkheimiana .
par=Agrofolha—94b-2: **Formada** pelos municípios de Lucas do Rio Verde, Tapurah, Nova Mutum e Sorriso, a região (a 350 km de Cuiabá) está produzindo 1,1 milhão de toneladas de soja em cerca de 400 mil hectares .
par=9369: **Decorado** de verde e amarelo e com preços aumentados em até 100 % -- o caldinho de feijão, por exemplo, subiu de R$ 1, na sexta, para R$ 2, na segunda --, o bar Coringa virou um pequeno Maracanã .
par=Agrofolha—94a-2: No total, **computadas** as coberturas e os animais, o faturamento bruto foi de R$ 338,7 mil .

Dos 368.108 particípios passados do corpo NILC/São Carlos, 76.404 funcionam como adjetivos pospostos e 13.831 como adjetivos antepostos. Além disso, 19.310 funcionam como nomes, e temos 130.868 orações participiais que seguem o conceito que modificam[5], e 14.472 que o precedem.

#8 As orações gerundivas são mais frequentes no português do Brasil?

Visto que uma das marcas diferenciais do português do Brasil em comparação com o de Portugal é o seu uso da progressiva com gerúndio, contrastando com a progressiva com *a*+ infinitivo, seria de esperar que as orações gerundivas também fossem significativamente mais usadas em português brasileiro. Usando o corpo CONDIV (Soares da Silva, 2008), contudo, vemos que tal fato não é incontroverso: a porcentagem é um pouco mais elevada, mas não decididamente: com efeito, 5,19 em 1000 palavras em português de Portugal contra 6,25 em português do Brasil são gerúndios numa oração gerundiva.

[5] Seguem: [temcagr=".*PCP.*" & func="ICL-(<.*|N<)"] e antecedem: [temcagr=".*PCP.*" & func="ICL-.*>"]

Se observarmos esta característica ao longo das três décadas contidas no corpo, respectivamente a década de 1950, a de 1970 e a de 2000, ilustradas na Tabela 4, vemos que, embora em português brasileiro a frequência seja sempre superior, a situação parece convergir para uma maior semelhança entre as duas variantes, ambas diminuindo o número de orações gerundivas.

Tabela 4 — A proporção de gerúndios de uma oração gerundiva por 1.000 palavras no corpo. CONDIV.

Variante	**ger.**	**Total**	**%**	**ger.**	**Total**	**%**	**ger.**	**Total**	**%**
	década de 1950			**década de 1970**			**década de 2000**		
PT	7225	1279115	5,6	6101	1204692	5,1	3921	837203	4,7
BR	5906	821426	7,2	4877	843826	5,8	5276	970864	5,4

Fonte: elaborada pela autora.

#9 Quais os gentílicos mais mencionados nos jornais brasileiros?

Considerando agora algo menos gramatical, podemos tentar identificar, nos jornais a que temos acesso, quais as palavras referentes a nacionalidade ou região mais usadas nos jornais do Brasil (nas décadas em que temos material: 1950, 1970, 1994-1995, 2000, 2010). Naturalmente que este é apenas um estudo exploratório, para demonstrar as capacidades da Gramateca, e não pretende apresentar resultados definitivos sobre a comunicação social ou o Brasil em geral. Também é um estudo que exemplifica uma das muitas informações fornecidas pelo PALAVRAS, o analisador sintático subjacente (BICK, 2000, 2007, 2014) [6].

Apresentamos os (primeiros) resultados na Tabela 5 como uma lista por ordem decrescente, e como uma nuvem de palavras na Figura 7, aproveitando para esclarecer que a *Folha de São Paulo* é uma das principais fontes do material jornalístico da Gramateca, o que explica naturalmente a grande quantidade de menções a *paulistas* e

[6] [classe="jorn" & sema=".*nat.*" & variante="BR"]

paulistanos. Convém também relembrar que as palavras *português* e *inglês* podem referir-se à nação ou à língua. Neste caso, as suas posições altas referem-se majoritariamente às línguas e não aos países[7].

Figura 7 — Nuvem de palavras gentílicas.

Fonte: elaborada pela autora.

[7] Existe além disso um clube de futebol chamado a *Portuguesa*, que contribui significativamente para a contagem: 9.105 dos 23.538 casos a ele se referem: 38,7%!

Tabela 5 — Gentílicos no texto jornalístico.

gentílico	**freq.**
brasileiro	96956
paulista	25903
português	23538
inglês	22307
índio	22157
norte-americano	21287
francês	20019
habitante	18681
alemão	15921
americano	13664
brasileira	13398
argentino	13035
italiano	12236
estrangeiro	11335
paulistano	10905
palestino	10320
guarani	10279
fluminense	9508
russo	9008
sérvio	8415
japonês	7767
carioca	7261

Fonte: elaborada pela autora.

5 Algumas explorações em outros corpos

#10 Quais as localizações mais frequentes no DHBB?

Finalmente, podemos aproveitar corpos específicos para fazer perguntas que só façam sentido nesse contexto. Já que temos o privilégio de disponibilizar o corpo Dicionário Histórico-Biográfico Brasileiro, uma obra que compreende muitos verbetes sobre personalidade e eventos relacionados com a história política moderna do Brasil, podemos, usando padrões de busca na Gramateca, identificar as relações familiares entre políticos (Higuchi et al., 2019), ou os locais de nascimento e de morte dos

verbetados, apresentados na Tabela 6. Em negrito se apresentam os locais onde mais políticos acabam a sua vida do que começam.

Tabela 6 — Locais de nascimento e de falecimento dos verbetados no DHBB.

Local	**Nascimentos**	**Mortes**
Rio de Janeiro	1039	1506
São Paulo	126	257
Recife	211	81
Salvador	201	81
Porto Alegre	163	86
Belo Horizonte	95	104
Fortaleza	131	43
Niterói	102	68
Curitiba	86	82
Campos	137	14
Belém	113	28
Brasília	6	122
Maceió	51	33
São Luís	65	19
Cuiabá	65	18
Petrópolis	50	32
Manaus	58	19
Natal	50	25
Aracaju	54	20
Teresina	45	28
Juiz de Fora	48	16
Florianópolis	34	26
Campinas	51	8
Goiânia	26	25
Pelotas	41	7
Pernambuco	41	5
Rio Grande do Sul	39	6
João Pessoa	24	19
Paris	7	29

Fonte: elaborada pela autora.

Embora evidentemente estejamos fazendo uma grande simplificação, podemos identificar os centros de poder político no Brasil observando as cidades em que muitos mais acabam a vida ao invés de lá nascerem: o Rio de Janeiro, São Paulo, Belo

Horizonte e Paris (!). Paris foi identificado por este método, mas claro que a conclusão neste caso tem de ser outra, não só por Paris não ser obviamente um centro do poder político brasileiro (e as mortes lá poderem se dever a um exílio, ou a um cargo diplomático), mas evidentemente porque muito poucos brasileiros nascem em Paris.

Em relação ao número de filhos[8] dos políticos, e não obstante a diferença entre os sexos em termos do conteúdo do DHBB ser muito marcada (só 204 verbetes se referem a mulheres, contra 6.457 sobre homens), verificamos que apenas 51% das mulheres verbetadas tinham filhos, contra 64% dos homens.

6 Considerações finais

Num volume sobre o tratamento computacional do português brasileiro, pode parecer estranho à primeira vista apresentar dois ambientes computacionais que se dedicam à língua portuguesa em todas as suas variedades, mas penso que ficou claro que essa é uma vantagem mesmo que o interesse primordial do pesquisador seja pela variante brasileira do português. Isto porque é possível selecionar só texto em português brasileiro, e ao mesmo tempo se pode comparar os resultados com mais material em português.

No artigo, escolhemos apresentar as potencialidades da Gramateca e da Literateca sobretudo a partir de exemplos de pesquisas relevantes para diferentes tipos de leitores, mas podemos também fazer uma sistematização aqui, até porque nem todas as possibilidades puderam ser exemplificadas. Mencionamos a criação de exercícios (enunciado e sua solução) para ensino da língua portuguesa, por meio do Ensinador, mas não a possibilidade de comparação entre duas buscas, para identificar diferenças entre distribuições, através do Comparador (este, e o Distribuidor, estão descritos em Simões e Santos (2014)). O Distribuidor faculta a obtenção de dados

[8] Para buscar o número de filhos no DHBB: ([word="uma?"]|[pos="NUM.*"]) [classe="biográfico" & lema="filh[oa]"]

extensivos em formato de planilha, para processamento subsequente em programas externos; se os dados forem apenas de distribuição (três colunas, portanto), podem se obter através da interface original, escolhendo-se a opção "Resultados em formato separado por ponto e vírgula". Ilustramos e comentamos a obtenção de nuvens de palavras como um resultado adicional, mas não referimos a criação automática de mapas, no caso de os corpos inquiridos terem geolocalização[9].

Além disso, não documentamos a ligação dos corpos com a possibilidade de criar formulários de consulta a informantes, por meio do Rêve, nem nos debruçamos sobre a criação de documentação de boas práticas de interligação com a linguagem R para a criação de outras visualizações mais complexas do material.

Independentemente de todas estas funcionalidades, é importante salientar que novas informações de anotação foram incluídas nos corpos (como os exemplos de emoção, etnicidade, relações familiares e geolocalização comprovam), e que, em última análise, é o aumento de informação, e a sua melhoria, que tornam estes ambientes úteis para a pesquisa.

Terminamos o artigo indicando que, apesar de termos, ao longo dos anos, desenvolvido ampla documentação e páginas de ajuda [veja-se por exemplo, FREITAS *et al*. (2011), em constante atualização], estamos sempre acessíveis para perguntas e pedidos de auxílio para permitir que todos possam fazer uso do material, assim como estamos abertos a sugestões de cooperação e de criação conjunta de novos recursos ou funcionalidades.

As páginas "oficiais" da Gramateca e da Literateca são respetivamente https://www.linguateca.pt/Gramateca/ e https://www.linguateca.pt/Literateca/,

[9] Esta funcionalidade não foi aqui ilustrada porque até agora apenas os corpos literários portugueses têm essa informação, devido ao financiamento do projeto BILLIG, mas a nossa intenção é continuar esse preenchimento para a parte brasileira num futuro próximo.

embora a interface principal continue a ser a do projeto AC/DC, https://www.linguateca.pt/ACDC/.

Agradecimentos

Agradeço sinceramente a Heliana Mello a sua gentil adaptação do texto para o português do Brasil.

Agradeço também a toda a equipe da Linguateca a existência do projeto e dos recursos, à FCCN – Fundação para a Computação Científica Nacional (Portugal), o alojamento da Linguateca nos seus servidores, e ao UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais.

Referências Bibliográficas

BICK, E. **The Parsing System "Palavras"**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Dissertação (PhD), Aarhus University. Aarhus University Press, 2000.

BICK, E. Automatic Semantic Role Annotation for Portuguese. *In*: **TIL**, V Workshop em Tecnologia da Informação e da Linguagem Humana. Rio de Janeiro, RJ, 30 de junho a 6 de julho de 2007. p. 1715-1719.

BICK, E. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. *In*: BERBER SARDINHA, T.; FERREIRA, T. L. S. B. (ed.). **Working with Portuguese Corpora**. London/New York: Bloomsbury Academic, 2014. p. 279-302.

FREITAS, C.; SANTOS, D.; GONÇALVES, A. **Perguntas já respondidas sobre o AC/DC**: desde como começar até uso complexo de funcionalidades poderosas. Primeira edição: 15 de Outubro de 2011. Disponível em: https://www.linguateca.pt/ACDC/

HIGUCHI, S.; SANTOS, D.; FREITAS, C.; RADEMAKER, A. Distant reading Brazilian history. *In*: NAVARRETA, C.; AGIRREZABAL, M.; MAEGARD, B. (ed.). **Proceedings of the Digital Humanities in the Nordic Countries 4th Conference** (Copenhagen, Denmark, March 5-8, 2019), 2019. p. 190-200.

LANGFELDT, M. C. Entre médicos e charlatães: A ascensão da medicina na formação da literatura brasileira. Apresentação no **III Encontro Nacional de Estudos Linguísticos e Literários (ENAELL), I Encontro Internacional de Pesquisas em Letras (ENIPEL)**. UEMA, 25-27 de maio de 2021. Disponível em: https://www.youtube.com/watch?v=XFEVaZC_ibU

SANTOS, D. Corporizando algumas questões. *In*: TAGNIN, S. E. O.; VALE, O. A. (org.). **Avanços da Lingüística de Corpus no Brasil**. Editora Humanitas/FFLCH/USP, São Paulo, 2008. p. 41-66. Disponível em https://www.linguateca.pt/Diana/download/Santos2008livroStellaOtofinal.pdf

SANTOS, D. Corpora at Linguateca: Vision and roads taken. *In*: BERBER SARDINHA, T.; FERREIRA, T. L. S. B. (ed.). **Working with Portuguese Corpora.** London/New York: Bloomsbury Academic, 2014. p. 219-236.

SANTOS, D. Gramateca: corpus-based grammar of Portuguese. *In*: BAPTISTA, J.; MAMEDE, N.; CANDEIAS, S.; PARABONI, I.; PARDO, T. A.S.; NUNES, M. G. V. (ed.). **Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014**, São Carlos/SP, Brazil, October 6-8, 2014, Proceedings. LNAI 8775. Heidelberg: Springer, 2014. p. 214-219.

SANTOS, D. Um novo corpo e seus desafios. *In*: FREITAS, C.; RADEMAKER, A. (ed.). **STIL 2015**: X Brazilian Symposium in Information and Human Language Technology and Collocated Events, Proceedings of the Conference, November 4 to 7, 2015. Natal, Rio Grande do Norte. p. 39-43.

SANTOS, D. José Mariano Gago - O Ministro da Língua. Apresentação no Encontro **Caminhos do conhecimento**, Leiria, 16 de Maio de 2018. Disponível em: https://www.linguateca.pt/Diana/download/LeiriaJMG.pdf

SANTOS, D. Doctors in lusophone literature. **Blog post in Digital Literary Stylistics** (SIG-DLS). 2019. Acessível de: https://dls.hypotheses.org/952

SANTOS, D. Explorando o vestuário na literatura em português. **TradTerm**, v. 37, n. 2, p. 622-643, 2021. DOI https://doi.org/10.11606/issn.2317-9511.v37p622-643

SANTOS, D.; ALVES, D. **Placing GIS and NLP in literary geography**: experiments with literature in Portuguese. Em apreciação. Disponível em: https://www.linguateca.pt/Diana/download/SantosAlves2022subm.pdf

SANTOS, D.; BICK, E. Providing Internet access to Portuguese corpora: the AC/DC project. In: GAVRILIDOU, M. et al. (ed.). **Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000.** Athens, 31 May-2 June 2000. p. 205-210.

SANTOS, D.; FREITAS, C. Estudando personagens na literatura lusófona. *In*: **STIL 2019** — XII Symposium in Information and Human Language Technology and Collocates Events, October 15-18, 2019, Salvador, BA, Proceedings of conference, 2019. p. 48-52.

SANTOS, D.; FREITAS, C.; BICK, E. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain. *In*: **OpenCor**, Canela, RGS, Brasil, 24 de setembro de 2018. Disponível em: https://www.linguateca.pt/Diana/download/CorLex.pdf

SANTOS, D.; MARQUES, R.; FREITAS, C.; SIMÕES, A.; MOTA, C. Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos. **Domínios de Linguagem,** v. 9, n. 2, abr./jun. 2015. p. 11-26. DOI https://doi.org/10.14393/DL18-v9n2a2015-2

SANTOS, D.; MOTA, C. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. *In*: CALZOLARI, N.; CHOUKRI, K.; MAEGARD, B.; MARIANI, J.; ODIJK, J.; PIPERIDIS, S.; ROSNER, M.; TAPIAS, D. (ed.). **Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010**, 17-23 May 2010, Valletta, Malta. European Language Resources Association, 2010. p. 1437-1444.

SANTOS, D.; RANCHHOD, E. Ambientes de processamento de corpora em português: Comparação entre dois sistemas. *In*: **Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR**, Évora, 20-21 de Setembro de 1999. p. 257-268.

SANTOS, D.; SIMÕES, A.; MOTA, C. Broad coverage emotion annotation. **Language Resources and Evaluation**, 2021. DOI https://doi.org/10.1007/s10579-021-09565-1

SIMÕES, A.; SANTOS, D. *Ensinador*: corpus-based Portuguese grammar exercises. **Procesamiento del Lenguaje Natural,** v. 47, septiembre de 2011, p. 301-309.

SIMÕES, A.; SANTOS, D. Nos bastidores da Gramateca: uma série de serviços. *In*: **Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish**, at PROPOR 2014, São Carlos, Brazil, 9 de outubro de 2014. p. 97-104.

SOARES DA SILVA, A. O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. In COSTA, L.; SANTOS, D.; CARDOSO, N. (org.). **Perspectivas sobre a Linguateca**: Actas do encontro Linguateca : 10 anos. Linguateca, 2008. p. 25-28. Disponível em: http://www.linguateca.pt/LivroL10/Cap04-Costaetal2008-Silva.pdf

Artigo recebido em: 31.10.2021 Artigo aprovado em: 02.05.2022