http://dx.doi.org/10.22409/gragoata.2017n44a1015

Estudos linguísticos e Humanidades digitais: corpus e descorporificação

Cláudia Freitas[a]

Resumo

Desde Saussure, a ciência linguística compartilha o pressuposto de que há uma língua/gem homogênea subjacente à variação. O presente artigo busca discutir este pressuposto, utilizando como metodologia explorações estatísticas em grandes corpora eletrônicos. Teorias são narrativas que tentam organizar os dados de que dispomos, sendo, portanto, razoável que novos dados (porque em quantidade muito maior, e em qualidade diferente) engendrem novas narrativas. Do ponto de vista apresentado aqui, vemos a língua ser regular e irregular, sem dentro ou fora; centro ou periferia. A irregularidade é incontornável; a linguagem é complicada e simples, simultaneamente.

Palavras-chave: *Humanidades digitais. Linguística pós-estruturalista. Corpus. Instabilidade da língua. Linguística computacional.*

Recebido em 31 de agosto de 2017
Aceito em 26 de dezembro de 2017

[a] Professora do Departamento de Letras da PUC-Rio/Programa de Pós-Graduação em Estudos da Linguagem; E-mail: claudiafreitas@puc-rio.br.

Introdução

Desde sua fundação como ciência autônoma, a linguística moderna tem como preocupação central a natureza da linguagem humana, em seus múltiplos aspectos – social-interacional; mental-biológico; textual-discursivo. Ainda que recortem objetos distintos, todos compartilham a busca pelo que seriam seus princípios constitutivos, subjacentes à inegável variação linguística.

Atribuímos a Saussure papel crucial nesse movimento, ainda que, como nos lembra PINTO (2008), a delimitação do objeto da linguística esteja completamente alinhada com o seu momento histórico: consolidação da ciência clássica; a relação entre língua e nação; e o conceito de indivíduo. "(..) uma língua de uma nação homogênea, falada por indivíduos homogêneos e integrados a essa nação, é um objeto privilegiado de uma nova ciência-piloto, pois apresenta-se como elemento positivo para ser verificado, analisado e sintetizado." (PINTO, 2008, p.1461). Imerso na tradição logocêntrica e engajado em conferir cientificidade à linguística, Saussure não escapa de um posicionamento que privilegia a busca pelo centro, pela estrutura da linguagem. Suas ideias (ainda que de maneira enviesada, apresentadas sob o ponto de vista dos editores do *Curso de Linguística Geral* (doravante *CLG*)), assim, dão ensejo aos principais direcionamentos nos estudos linguísticos, nos quais a dimensão estrutural/universal/sistêmica da linguagem se destaca. A fim de viabilizar sua empreitada epistemológica, Saussure faz uso de certas estratégias, uma delas a conhecida separação entre *langue* e *parole*; *linguística interna* e *linguística externa*. A palavra estratégia, não nos esqueçamos, remete à arte de aplicar com eficácia os recursos de que se dispõe, visando a alcançar determinados objetivos.

Neste artigo, proponho a utilização de uma outra estratégia, associada a um outro ponto de vista: observar a língua a partir da lente de ferramentas computacionais, tomando como objeto grandes corpora eletrônicos. Tal estratégia só é possível porque dispomos, hoje, de meios capazes de efetuá-la: computadores e um texto descorporificado. A utilização de computadores na área das humanidades já é uma realidade e a abordagem tem um nome específico: Humanidades Digitais. O impacto desse tipo de enfoque nos

estudos linguísticos, no entanto, ainda é bastante modesto, limitando-se, na maioria das vezes, à identificação e exploração de padrões léxico-gramaticais e, de maneira ainda mais tímida, à busca de exemplos para corroborar teorias já existentes. Mas, se o ponto de vista determina o objeto, que objeto Saussure veria se tivesse à disposição as tecnologias de hoje? Sabendo que teoria e dados estão em relação dialógica, é razoável imaginar que novos dados sobre a linguagem espontânea, materializados em grandes corpora, acarretem impactos para teorias linguísticas. Advogar em favor de uma visão de língua que ainda comparece pouco nos estudos linguísticos, sobretudo aqueles que se interessam por questões de natureza gramatical, é o principal objetivo do presente artigo.

Leituras não lineares e o texto descorporificado

Leitura é palavra polissêmica. Um de seus sentidos pressupõe a existência de um texto em sentido estrito – linguagem verbal. Nesse recorte, a leitura está associada a um certo tipo de atividade, que envolve o texto e um movimento de decodificação de sinais gráficos. Desde há algumas décadas, esse objeto tem aparecido com uma nova materialidade: no texto eletrônico, temos um texto visível na tela/monitor, mas que não *está* na tela, nem está no computador/microprocessador, do mesmo modo como um texto *está* em um livro ou revista. O texto eletrônico pode ser lido da maneira tradicional, mas essa outra – nova – materialidade também possibilita novos rearranjos. Para fazer referência a essa outra materialidade, tomo emprestado de Paixão de Souza (2013) o termo *descorporificação*. O texto digital é descorporificado porque não tem corpo, não tem um suporte específico, está em lugar nenhum: a nova materialidade descorporifica. O documento digital é um objeto virtual ao qual atribuímos sentido, não é (mais) um objeto físico. O que lemos na tela é um simulacro de texto: letras, disposição espacial, paragrafação e margens não passam de instruções que serão seguidas por programas, e que nos farão ver páginas, margens e linhas. O texto digital é um aglomerado de codificações, e tal fato carrega uma série de implicações. A seguir trago alguns exemplos para ilustrar a questão.

Elaborada a partir de um conjunto de 57 obras da literatura brasileira do século XIX (2.6 milhões de palavras, 19 autores diferentes), a figura 1 apresenta a distribuição dos predicadores humanos mais frequentes por gênero do personagem. O material (em constante crescimento) refere-se ao corpus OBras, que contém obras da literatura brasileira em domínio público (a documentação detalhada do material encontra-se na página do projeto[1]). Para produzir a lista, foram feitas consultas linguísticas, por meio da interface do serviço AC/DC (COSTA et al., 2009), capazes de indicar a presença de predicadores humanos por gênero, como a busca por estruturas predicativas ou apositivas[2]:

— Sempre, Leopoldo, sempre **ela é bela**, **formosa**, **encantadora**, **angélica**! (*A Moreninha*, Joaquim Manuel de Macedo)

— Uma *moça bonita*, que podia fazer um casamento importante... (*Turbilhão*, Coelho Neto)

— A **Viscondessa do Rio Seco**, **trigueira** e **fina**, exibe, com vaidade, o seu riquíssimo vestido chegado de Paris. (*A Marquesa de Santos*, Paulo Setúbal)

— É um **homem sério** e **destemido**! (*O Cortiço*, Aluísio Azevedo)

Figura 1. Predicadores humanos por gênero em 57 obras da literatura brasileira do sec. XIX. Nos predicadores masculinos "fazer" corresponde à homem *feito*.

[1] http://www.linguateca.pt/OBRAS/OBRAS.html

[2] Os dados foram obtidos por meio de consultas como [pos="PROP" & func="SUBJ>"] [lema="ser|estar"] [pos="ADV.*"]* @ [temcagr!=".*PASS.*" & pos="ADJ|N|V" & gen="F" & func="<SC"] na interface do AC/DC http://www.linguateca.pt/ACDC/

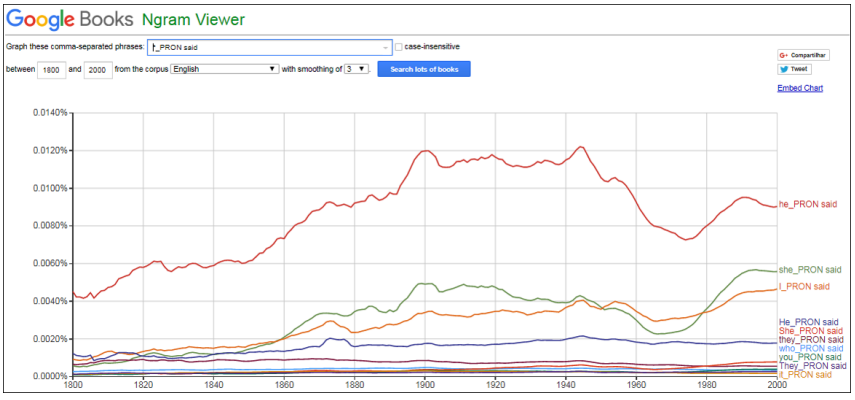
Quanto aos dados propriamente, as figuras apresentam os 20 predicadores mais frequentes para cada gênero, e o tamanho é proporcional à frequência. Se tomamos apenas os três predicadores mais frequentes de cada gênero, percebemos

que, quando se trata de caracterizar personagens femininos, o corpo (ou, características físicas) são as que mais se sobressaem: são bonitas, loiras e belas. Quanto aos personagens masculinos, as caracterizações são de caráter ou posição social: são bons, sérios e ricos. Embora a figura dê margem a uma análise mais elaborada, a ideia aqui é apenas ilustrar uma aplicação.

Considerando apenas os predicadores femininos (em menor quantidade que os masculinos; há menos personagens femininos), as buscas retornaram um total de 943 predicadores, distribuídos em 492 predicadores diferentes. A geração das listas levou alguns minutos, e não é difícil perceber o potencial desse tipo de análise.

Na figura 2, apresento os resultados de uma consulta simples utilizando o recurso *N-gram Viewer*, do Google[3], que realiza buscas em uma base de livros publicados entre os anos 1800 e 2000. A consulta envolve a visibilidade de vozes: como é a distribuição de falas (nesse material) no que se refere ao gênero? Em alguns segundos, vasculhamos todo o conteúdo e vemos a incidência muito maior de "*He said*" ("Ele disse") sobre "*She said*" ("Ela disse"). O ponto – nas figuras 1 e 2 – é que este tipo de análise só é possível quando temos o texto *descorporificado*.

Figura 2. Distribuição de pronomes pessoais que atuam como sujeitos de "*said*" na base de livros do Google



Trata-se de ler o texto de outra maneira, não linear. Nos estudos literários, já se cristalizou o termo *distant reading* (MORETTI, 2000) para fazer referência a esse tipo de leitura. Não criamos um texto novo, no sentido convencional de "novidade", mas o observamos de

[3] https://books.google.com/ngrams

um ponto de vista diferente, o que nos permite *ver* coisas diferentes. Os dados não são novos, mas até algum tempo atrás, a forma de obtê-los seria ler e destacar, caso a caso, cada uma das caracterizações e organizá-las de alguma maneira. Essa estratégia convencional de leitura esbarra em dois obstáculos, a princípio intransponíveis: (a) nosso tempo de leitura e de preparação do material, algo talvez intangível no tempo de trabalho de um projeto de pesquisa; (b) a análise exigida, uma análise minuciosa, não é compatível com uma leitura em larga escala.

Esse tipo de pesquisa integra o vasto campo das Humanidades Digitais (HD), uma abordagem para todo o campo das Humanidades que faz uso intensivo de recursos e ferramentas digitais. As HDs têm como princípios o uso de dados abertos, o compartilhamento e uma interdisciplinaridade profunda. No âmbito das HDs que têm foco na exploração automática de grandes acervos textuais – ficcionais ou não –, o foco não está em análises estritamente linguísticas, ainda que estas sejam enriquecedoras para boa parte dos trabalhos.

Volto às figuras 1 e 2: mesmo simples, as buscas que dão origem aos dados fazem uso de ferramentas desenvolvidas especialmente para lidar com uma língua – ferramentas que identificam palavras e suas propriedades morfossintáticas, e realizam análises sintáticas, por exemplo. Se, no contexto das HDs, a combinação de textos e ferramentas tem trazido contribuições valiosas, de novas respostas para antigas perguntas, e mesmo para a elaboração de novas perguntas, é razoável imaginar que essa nova maneira de observar a língua possa também impactar estudos linguísticos.

Implicações do texto descorporificado para os estudos linguísticos

Desde os trabalhos de Biber (1985) e Sinclair (1983;1991), a articulação entre grandes corpora eletrônicos e computadores abriu uma larga avenida para a exploração da língua de um ponto de vista quantitativo, viabilizando sobretudo o estudo de gêneros textuais e a depreensão de padrões léxico-gramaticais. Mais recentemente, a pesquisa linguística com corpus[4] já dispõe de uma série de ferramentas e serviços desenvolvidos especialmente para o estudo da língua que permitem ir além da busca por padrões (veja-se, por exemplo, SANTOS (2011; 2014) para a apresentação de alguns recursos e ferramentas públicos para a língua portuguesa).

[4] Lembro, com Leech (1992) e Santos (2008), que um corpus não é objeto de estudo de um linguista; o corpus é apenas um meio pelo qual observamos materializações de uma língua, e por isso pode ser combinado com os diversos ramos da pesquisa linguística. Por isso, também, como defendido por Santos (2008), a pouca precisão da tradução Linguística de Corpus para Corpus Linguistics, ainda que esta já esteja cristalizada em português. Afinal, não se diz "Linguística de Sala de Aula" para fazer referência a estudos linguísticos que tomam por base as interações verbais em sala de aula, por exemplo, e por isso a ideia de uma linguística com corpus talvez seja mais adequada.

Nesse contexto, o papel do linguista é reformulado. Segundo de Beaugrande (2002),

> Corpus research recasts the linguist: not in the role of the "ideal speaker-hearer in a completely homogeneous speech-community, who knows its language perfectly", but in the role of an ordinary speaker-hearer (and writer-reader) in a heterogeneous community, who knows its language only partially and actively seeks access to the knowledge of others. We claim authority for our statements not from harbouring super-human powers of introspection (1.9), but from examining large sets of authentic data. (de BEAUGRANDE, 2002:119)

No entanto, a reformulação do papel do linguista é pouco. Os exemplos apresentados evidenciam também o potencial das ferramentas – lentes que ampliam nossa visão. Anthony (2013) argumenta que, para a linguística, a relevância das ferramentas compara-se à relevância do microscópio e do telescópio para a biologia e a astronomia, respectivamente. Como nos lembra Borges Neto (2011), teorias são sempre uma tentativa de explicação a partir de dados, e dados são, simultaneamente, a motivação e a exemplificação de uma teoria. Com alteração nos dados, em quantidade e qualidade, é razoável que teorias sejam repensadas.

O que viu Saussure, o que vê a ciência linguística

Desde Saussure e a bem sucedida empreitada estruturalista de tornar a linguística uma ciência-piloto exemplar[5], a dimensão regular/homogênea/estrutural/sistêmica da língua tem sido privilegiada (refiro-me sobretudo às áreas dos estudos da linguagem que não sofreram influência das reflexões associadas à Desconstrução). Para compreender e/ou descrever *como a linguagem é*, é necessário distinguir entre o essencial e o acidental; para depreender sua(s) estrutura(s), (ou, de forma atualizada segundo a escola gerativa, os *princípios e parâmetros*), é preciso se despojar de sua visível variabilidade, responsável por inviabilizar generalizações. Saussure propõe, como uma das tarefas da Linguística

> "procurar as forças que estão em jogo, *de modo permanente e universal*, em todas as línguas, e deduzir as *leis gerais* às quais se possam referir todos os fenômenos peculiares da história". (*CLG*, p.13, grifo meu)

[5] Não se pode negar que o custo dessa empreitada foi bastante alto, tendo como uma das consequências o distanciamento da linguística da vida real, como abordado, por exemplo, em Pennycook (2004).

e reconhece a heterogeneidade da linguagem – "aglomerado confuso de coisas heteróclitas, sem liame entre si" (*CLG*, p.16), bem como a ambição da empreitada – "em nenhuma parte se nos oferece integral o objeto da Linguística" (*CLG*, p.16) – e assume, como estratégia, o recorte da *langue*, estrutura organizacional associada à estabilidade:

> "Há, nos parece, uma solução para todas essas dificuldades: é necessário colocar-se primeiramente no terreno da língua e tomá-la como norma de todas as outras manifestações da linguagem." (*CLG*, p.16)

A *langue* é "social em sua essência e independente de indivíduo" (*CLG*, p.27), fruto da eliminação de tudo o que é externo ao sistema, e se opõe à *parole*, individual e acessória (*CLG*, p.22). O projeto de uma ciência linguística geral se funda, assim, na oposição regular/irregular; homogêneo/heterogêneo; interno/externo, cabendo sempre ao primeiro elemento de cada par a posição de objeto central e legítimo. Vemos em Mattoso Camara: "De qualquer maneira, a invariabilidade profunda, em meio a variabilidades superficiais, é inegável nas línguas" (CÂMARA JR., 1982, p.17). E mesmo quando o objeto de estudo é a pluralidade das manifestações da língua em uso, a observação da variabilidade tem, como objetivo último, a depreensão das estruturas ou sistemas subjacentes à tal variação (veja-se, por exemplo, a proposta de *frames* da linguística cognitiva ou as metafunções da linguística sistêmico-funcional). A visão de língua como composta por um sistema de regras com grau de estabilidade interna (e, do mesmo modo, por exceções) não é alvo de críticas diretas, como o são (ou podem ser) a questão da autonomia linguística, a dicotomia interno/externo e o caráter positivo do signo.

Não podemos esquecer, no entanto, que a homogeneidade e a estabilidade são apenas presumidas; são hipóteses de trabalho. Trata-se, assim, de uma tentativa de explicação sobre a linguagem e, enquanto tal, deveria poder ser verificada empiricamente, já que "a língua é objeto de natureza concreta, o que oferece grande vantagem para o seu estudo" (*CLG*, p.23).

Apesar disso, grande maioria dos trabalhos linguísticos com base em corpus não questiona os pressupostos saussurianos quanto à existência de um caráter nuclearmente regular/homogêneo/estruturado da língua, alinhando-se, ainda que

guardem uma miríade de diferenças entre si, no que podemos chamar de perspectiva logocêntrica.

Como apontado em Santos (2017), abordagens estatísticas foram, até o presente momento, pouco referidas em estudos linguísticos, embora a situação já comece a mudar: "While in the '90s most corpus linguists ignored statistical methods, now it is almost impossible not to apply them and get published" (2017). No entanto, é a mesma autora que nota, por outro lado, que uma propriedade das línguas humanas, capturada pela chamada "lei de Zipf" tem suas consequências pouco exploradas no contexto dos estudos linguísticos (SANTOS, 2008; 2014). Manning & Schutze (1999), no contexto mais informático do processamento (automático) de linguagem natural (PLN), também defendem os benefícios de uma abordagem estatística para lidar com a língua. A seguir, explorarei algumas propriedades estatísticas da língua, argumentando em favor de sua absoluta relevância para um outro olhar no contexto de estudos linguísticos contemporâneos.

Insights estatísticos e seu impacto em estudos linguísticos

Estatística pressupõe contagem, que por sua vez pressupõe estabilidade – ou consenso – quanto às unidades que serão contadas. No caso da língua, como sabemos, essas unidades não estão previamente dadas, e a pluralidade de discussões em torno do construto teórico *palavra* é uma boa ilustração deste ponto – e este é apenas um alerta contra a incorporação ingênua de contagens sobre a língua. Ainda assim, e de maneira simplificada, ao observar a distribuição das palavras em um texto, encontramos variabilidade e heterogeneidade, o que em nada surpreende, pelo contrário: é justamente devido à variação que boa parte dos estudos linguísticos opta por recortar, como objeto de pesquisa, uma língua homogênea.

No entanto, a referida heterogeneidade pode esconder sistematicidades surpreendentes[6]. Tomemos como exemplo *Dom Casmurro*. O livro tem um total de 65.639 palavras, distribuídas em 5.994 palavras diferentes[7]. A figura 3a mostra a distribuição das 30 palavras (lemas) mais frequentes, em

[6] O conteúdo estatístico apresentado aqui é bastante superficial, mas espero suficiente para ilustrar o ponto em questão. Me utilizo, sobretudo, de Santos (2008 e 2011) e Manning & Schutze (1999).

[7] A definição de palavra, como sabemos, não é consensual. Levando em conta o material consultado aqui, indico que itens como raio-X, DOI-Codi, vice-reitor e Rio de Janeiro são considerados uma unidade de análise.

ordem decrescente, indicando o número de ocorrências de cada uma. Concluímos daí que as 10 palavras mais frequentes do livro correspondem a 29% de todas as palavras do livro. Já a figura 3b apresenta, para o mesmo livro, a frequência das frequências, considerando desde palavras que aparecem 7 ou menos vezes até palavras que ocorrem 100 ou mais vezes. Dela, temos que as palavras raras no livro (com frequência menor ou igual a 7) correspondem a cerca de 87.5% de todas as palavras do livro e, por outro lado, que as palavras mais comuns (que aparecem 100 ou mais vezes) respondem por apenas 1.4% das palavras do livro.[8] Tomando apenas a classe dos verbos de *Dom Casmurro*, temos uma situação parecida, como ilustra a figura 3c. Pouquíssimos verbos – apenas 19 verbos, que não chegam a 2% de todos os verbos do livro – aparecem mais de 100 vezes e, por outro lado, verbos que aparecem pouco, com até 5 ocorrências apenas, respondem por 75% de todos os verbos da obra. Temos, ainda, que quase 40% de todos os verbos (429 verbos) aparecem apenas uma vez – *hapax legomena*. À primeira vista, podemos imaginar que tal variedade lexical é uma característica do texto literário, mas as figuras 4a e 4b apresentam o mesmo tipo de distribuição de palavras, tendo por base textos de outra natureza: entrevistas transcritas que versam sobre a vida dos entrevistados (corpus Museu da Pessoa), textos de jornal (corpus CHAVE) e o OBras, já mencionado (e que contém *Dom Casmurro*)[9]. Como é possível observar, a distribuição se repete, independentemente do tipo de texto.

Figura 3a. Lista das 30 palavras mais frequentes em *Dom Casmurro*, em ordem decrescente de frequência

1.	o	4334	11.	em+o	758	21.	ter	421
2.	que	2548	12.	a	653	22.	dizer	385
3.	e	2167	13.	se	601	23.	outro	372
4.	ser	1902	14.	meu	598	24.	como	368
5.	de	1867	15.	mas	568	25.	por	345
6.	não	1499	16.	ele	559	26.	Capitu	330
7.	de+o	1493	17.	para	537	27.	capitular	293
8.	um	1183	18.	com	534	28.	José Dias	292
9.	eu	1180	19.	ir	468	29.	estar	290
10.	a+o	758	20.	em	442	30.	poder	287

[8] Esta apresentação estatística se inspira na apresentação que Manning e Schutze (1999) fazem de Tom Sayer.

[9] Todo o material integra o acervo do AC/DC.

Figura 3b. Distribuição da frequência das frequências de todas as palavras em *Dom Casmurro*

Frequência	Frequência da frequência
<7	5088
10	69
11-50	503
51-100	68
>100	85

Figura 3c. Distribuição da frequência das frequências de todos os verbos em *Dom Casmurro*[10]

Frequência	Frequência da frequência
1	429
2	192
3	80
4	53
5	32
6	38
7	34
8	21
9	12
10	10
11-50	117
51-100	14
> 100	19
Total	1051

[10] Dados obtidos com a expressão [autor="MdA"& obra="Dom.*" & pos!="PU"& pos="V"]

Figura 4b. Distribuição da frequência das frequências em 3 materiais distintos

Frequência	Frequência da frequência		
	Museu da Pessoa (1.8 milhão de unidades*)	OBRAS (1.6 milhão de unidades)	CHAVE (124.1 milhão de unidades)
1	11167	15828	569308
2	3630	4634	134525
3	1863	2772	55683
4	1186	1863	33402
5	836	1355	21071
6	595	1116	14977
7	485	845	11278
8	400	744	9031
9	344	595	7295
10	328	522	6024
50	22	28	411
51-99	663	969	11946
> 100	937	1041	23774
Total	22.456	32.312	898.725

* Consideram-se sinais de pontuação e outros símbolos, como $ e %, por exemplo.

Figura 4a. 20 unidades mais frequentes em cada um dos 3 corpora, em ordem decrescente de frequência

Posição	Museu da Pessoa		OBRAS		CHAVE	
1	o	133309	o	141879	o	13458492
2	,	119156	,	112401	de	8093504
3	.	70864	de	79123	,	6741178
4	de	68021	.	48067	.	3908180
5	ser	46558	e	35302	em	3085083
6	que	45678	que	30045	que	2222459
7	em	37778	em	28027	e	2131199
8	e	36961	a	23359	a	1979652
9	eu	34414	um	22785	ser	1873195
10	um	33092	ser	18954	um	1691086
11	ter	25123	ele	16055	por	1081492
12	não	20482	não	14463	«	1066141
13	a	18049	;	12349	»	1015144
14	para	15764	--	11181	para	951469
15	?	15482	por	10946	com	784114
16	--	14801	!	10930	não	737172
19	ir	13594	com	10787	ter	602757
18	com	12689	se	9954	se	581601
19	estar	12314	para	9056	)	574479
20	muito	12022	eu	8913	(	563791

A figura 4b nos mostra que palavras raras, com frequência de até 10 vezes em cada um dos materiais, são a maioria na língua, e correspondem a 93% de todas as palavras (Museu da Pessoa e OBras) e 96% (CHAVE). Por outro lado, palavras frequentes, que aparecem 100 ou mais vezes, correspondem a 4.1%, 3.2% e 2.6%, no Museu da Pessoa, OBras e CHAVE, respectivamente. Por outro lado ainda, levando em conta apenas as palavras que aparecem com mais frequência (figura 4a), vemos que, em cada um dos 3 corpora, as 10 unidades mais frequentes correspondem a 35% (Museu da Pessoa), 33% (OBras), 36% (CHAVE) de todas as palavras do material. Uma última observação diz respeito à proximidade da distribuição dos números, independentemente do tipo de texto e do tamanho do material analisado[11].

Esse fenômeno relativo à distribuição das palavras em um texto foi capturado por Zipf[12] e, como mencionado, não costuma ser levado em conta em estudos linguísticos que fazem uso de grandes corpora. Um ponto associado à chamada lei de Zipf é que a quantidade de casos raros se mantém mesmo quando ampliamos a nossa mostra. Ou seja, a "cauda longa" (o grande número de casos com pouca frequência e que aumenta à medida que a frequência diminui) na distribuição da frequência não é decorrente de uma amostra insuficiente; é, antes, característica da língua – quase todas as palavras são raras[13].

Outro ponto observável a partir dos números é a existência de um padrão na distribuição: pouquíssimos casos com muitas ocorrências, um número intermediário de frequência média, e um número enorme de casos de frequência baixa. Além disso, e mais importante, essa regularidade na distribuição não se aplica apenas a palavras. Santos (2008) ilustra o fenômeno tomando como unidade de análise a tradução do tempo verbal entre português e inglês. De maneira complementar, tomo como unidade de análise a classe de formas dos sintagmas nominais. Considerando o corpus Bosque-UD (RADEMAKER et al., 2017), temos um total de 41.366 SNs, espalhados em 11.613 estruturas distintas[14]. As figuras 5a e 5b são complementares e dão uma ideia da distribuição dos SN pelo corpus, composto por textos jornalísticos. Os números seguem a mesma linha das análises anteriores: estruturas raras (com frequência >10) correspondem a 98% de todas as estruturas. Por outro lado,

[11] Todos os números são públicos para a consulta, na página do Ordenador, serviço associado ao AC/DC.

[12] Formulada em 1945, a chamada Lei de Zipf é um grupo de fenômenos (MANNING e SCHUTZE, 1999)

[13] Manning e Schutze (1999) observam que a relação entre frequência e posição parece ter sido notada pela primeira vez por Estoup, ainda em 1916, mas foi Zipf quem a popularizou.

[14] Importante notar que a anotação com base em modelo de dependências sintáticas não contém sintagmas nominais. Para os dados apresentados aqui foram considerados SN todos os elementos que são dependentes de um substantivo. Além disso, uma especificidade da anotação de Universal Dependencies determina que preposições e conjunções sejam dependentes dos substantivos que os seguem. Por isso, alguns ajustes precisaram ser feitos para capturar a noção de SN. Agradeço a Alexandre Rademaker pela geração dos dados a partir do Bosque.

apenas 1 estrutura (DET NOUN), a mais frequente, dá conta de 20% de todas os casos de SN e, se considerarmos as 5 estruturas mais frequentes, temos quase metade (43%) de todos os SNs. A figura 5b explicita a enorme quantidade de casos (mais de 10 mil) com apenas uma ocorrência[15], em contraste com casos com mais de 100 ocorrências (26). A figura 5a lista as 20 estruturas mais frequentes, em ordem decrescente. Tomados em conjunto, esses dados põem em xeque a ideia de uma língua homogênea subjacente à variação, alvo da descrição e investigação linguística.

Figura 5a. Lista com as 20 estruturas mais frequentes de SN no Bosque-UD, em ordem decrescente de frequência

Estrutura	Frequência	Estrutura	Frequência
DET NOUN	8363	NOUN ADP DET NOUN	381
NOUN	5699	DET NOUN ADP DET PROPN	332
DET NOUN ADJ	1644	DET NOUN ADP PROPN	251
DET NOUN ADP DET NOUN	1185	ADJ NOUN	250
NOUN ADJ	924	DET NOUN ADP DET NOUN ADJ	203
NUM NOUN	738	NOUN CCONJ NOUN	173
DET NOUN ADP NOUN	640	DET NOUN PROPN PROPN	166
DET ADJ NOUN	602	DET NOUN PROPN	152
DET DET NOUN	564	NOUN ADP DET PROPN	151
NOUN ADP NOUN	520	DET NOUN ADJ ADP DET NOUN	136

Figura 5b. Frequência das frequências do SN no Bosque-UD

Frequência	Frequência da frequência
1	10515
2	496
3	165
4	84
5	63
6	38
7	34
8	21
9	18
10	14
11-50	121
51-100	18
> 100	26

[15] Ainda que o material tenha passado por uma revisão linguística, sabemos que é praticamente impossível que um corpus deste tamanho não contenha análises com erros, e acreditamos que tais erros devem aumentar a quantidade de casos com apenas uma ocorrência.

Um outro ponto de vista

Quando tomamos a perspectiva estatística, damo-nos conta de que fenômenos que ocorrem muito pouco constituem uma vasta porção da língua, fragilizando o pressuposto de que as irregularidades são fatos de superfície. Daí que, quando propomos que à linguística deve caber, centralmente, o estudo do que é regular/homogêneo/universal, nos esquecemos de que a aposta sobre a existência de tal homogeneidade é nada mais do que uma aposta.

Na língua, o que não se enquadra como padrão não necessariamente se enquadra como fenômeno periférico. Se, por um lado, tratar a língua somente sob a perspectiva da irregularidade é inevitavelmente pintar um quadro bastante distorcido do que seja uma língua, deixando uma série de fenômenos de fora, por outro lado, tratar a língua somente sob a perspectiva da regularidade tem como consequência a mesma distorção.

O que tentei apresentar nas seções anteriores foi uma outra estratégia para lidar com a variabilidade: vastas porções da língua em uso. Em ambos os casos – língua homogênea e corpus – estamos diante de substitutos para lidar com um todo que nos é inacessível. No entanto, acredito que o exame de grandes porções da língua em uso é uma estratégia que oferece vantagens quando o foco está na realização de estudos sistemáticos. Por um lado, grandes corpora podem ser capazes de refletir a língua de uma coletividade por meio da soma das individualidades, característica valorizada por Saussure e atribuída à *langue;* por outro, a valorização da linguagem em uso não deixa de ser também incentivo para aproximação de trabalhos com grandes corpora eletrônicos; por outro lado ainda, de um ponto de vista metodológico, o trabalho com corpus oferece não apenas sustentação empírica para o estudo de um objeto que é concreto, mas também um ambiente que permite a reprodução de resultados, desde que os estudos sejam feitos com corpora públicos, com dados públicos e compartilháveis, como os utilizados aqui.

Ainda de um ponto de vista metodológico, espera-se da ciência linguística, "naturalmente" enquadrada no paradigma lógico-científico, logocêntrico, que apresente (i) um objeto estável; (ii) exaustividade descritiva; (iii) poder de generalização (das observações).

O primeiro ponto, estabilidade do objeto, busca assegurar a não-contradição como princípio mais básico: o objeto de pesquisa não pode oscilar, não é possível (segundo

a opção pelo logocentrismo) ser e não-ser, simultaneamente. Porém, o que as explorações estatísticas nos obrigam a ver é um *objeto oscilante*: regular e irregular, simultaneamente. Se consideramos apenas as figuras 3a, 4a, 5a, temos um quadro que evidencia a homogeneidade: poucas unidades ou construções são muitíssimo frequentes. Por outro lado (figuras 3b, 4b, 5b), vemos que construções raras, quando somadas, constituem uma enorme parcela do todo. Quando tratamos de exceções, fenômenos raros ou periféricos, estes só são assim considerados se tomamos a frequência de cada um, isoladamente. Considerados como um todo, fenômenos periféricos são centrais.

A exaustividade descritiva e o poder de generalização também ficam debilitados quando levamos em conta as análises estatísticas. Com tantos casos que ocorrem apenas uma vez, e já sabendo que não adianta ampliar a amostra se o objetivo é ter mais dados para forçar a generalização, lidar com a ideia de exaustividade será frustrante. O que não significa o abandono da generalização, apenas o reconhecimento de suas limitações.

A crítica de Derrida ao estruturalismo de Saussure centra-se, sobretudo, na caracterização do signo linguístico, que surge como entidade positiva a partir da união de entidades que são nada mais que diferenças (DERRIDA, 1973). A aposta saussuriana em uma língua homogênea não está especialmente em foco, mas é atingida pela desconstrução de todo o empreendimento logocêntrico sustentado na estabilidade do objeto. A partir do ponto de vista explorado aqui, conseguimos, de alguma maneira, capturar o dinamismo-instabilidade da língua, levá-lo em consideração, ao invés de criar estratégias que justifiquem escamoteá-lo, sem precisar deixar de lado o que se repete na linguagem/nos usos. Vemos a língua ser regular e irregular, sem dentro ou fora; centro ou periferia. A irregularidade é incontornável, assim como a regularidade é necessária; a linguagem é complicada e simples, simultaneamente.

Considerações finais

Ao longo deste artigo, procurei mostrar como algumas ideias da estatística linguística podem oferecer subsídios para um tipo de pensamento sobre a linguagem que ainda

é raro em estudos linguísticos, principalmente aqueles mais frequentemente associados aos conteúdos gramaticais. O ponto de vista sugerido aqui não é novo – pelo contrário, a ideia de instabilidade na linguagem e a sua resistência a teorizações generalizadoras são alvo do Wittgenstein das *Investigações Filosóficas* (1953), e tem raízes que remontam ao pensamento sofista. Mas temos elementos capazes de enriquecer e ilustrar a discussão, tirando proveito da concretude descorporificada da linguagem. Com isso, abordagens linguísticas que normalmente dialogam pouco com uma visão não-logocêntrica, por terem como interesse objetos em geral pouco tematizados por estas abordagens – como morfologia e sintaxe – podem se beneficiar da fartura de dados, por um lado, e de alguns direcionamentos sobre como enfrentá-los, por outro, levando em conta a variedade e imprevisibilidade, e não os deslocando para os espaços de interface ou de exceção.

A dinamicidade da linguagem é reiteradamente reconhecida por Saussure, ainda que seja identificada como fenômeno acidental. No entanto, o reconhecimento da centralidade da instabilidade (ou dinamicidade) da língua não inviabiliza a realização de estudos sistemáticos. Antes, abre caminho para outros tipos de estudos linguísticos, que podem abraçar a instabilidade (dinamicidade) sem abrir mão de sistematicidade. Traz materialidade ao que se nos apresenta como intangível.

Este posicionamento tem uma série de consequências. Do ponto de vista didático, por exemplo, confronta a forma como certos conteúdos são tratados em manuais introdutórios de linguística, com a apresentação de uma língua essencialmente regular. Do ponto de vista de colaborações acadêmicas com outros campos, como o PLN, acarreta a pouca participação linguística. Trata-se de uma área dedicada à resolução de problemas que têm a linguagem como insumo principal (e o desenvolvimento de ferramentas para estudos linguísticos é, infelizmente, sua dimensão menos valorizada), mas que dialoga de maneira escassa com a linguística, sobretudo porque as contribuições mais substanciais que esta teria a oferecer se baseiam em uma língua idealizada (veja-se SAMPSON, 2003). Os dados subjacentes à figura 2, por exemplo, só existem porque, para todo o conteúdo dos livros consultados, foi feita uma análise gramatical que nos permite

buscar por itens metalinguísticos como *pronome*. Para que a categoria *pronome* seja atribuída a uma palavra, é preciso ter claro – ou, pelo menos, é preciso consenso sobre – o que deve contar como um *pronome*. O mesmo vale para *substantivos*, *adjetivos* e demais itens metalinguísticos – e a discussão acerca de tais classificações, alvos recorrentes de disputas nos estudos gramaticais, tem voltado à tona com as necessidades do processamento automático[16]. Quando a língua em uso é tomada como objeto e quando entendemos as limitações de abordagens que consideram apenas a regularidade, temos mais chances de oferecer contribuições linguísticas relevantes para o PLN, superando assim o isolamento entre ambas as áreas, já muito bem apontado por Sparck-Jones (2007).

Dar conta de ambas as dimensões - regularidade e irregularidade -, simultaneamente, é uma tarefa ambiciosa, e nem se espera de um pesquisador que seja capaz disso. Mas a consciência de que essas dimensões são igualmente "língua" ajuda a lidar com as exceções, diminui expectativas e frustrações.

REFERÊNCIAS

ANTHONY, L. A critical look at software tools in corpus linguistics. *Linguistic Research* 30(2), 2013, p. 141-161. Disponível em: <http://isli.khu.ac.kr/journal/content/data/30_2/1.pdf>. Acessado em: 30 out. 2017.

BIBER, D. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics* 23 (2), 1985, p. 337–360.

BORGES NETO, J. Morfologia: Conceitos e Métodos. In: LIMA, M. A. F.; ALVES FILHO, F.; COSTA, C. S. C. (Org.). *Colóquios linguísticos e literários: enfoques epistemológicos, metodológicos e descritivos*. Teresina: Edufpi, 2011. p. 53-72.

COSTA, L.; SANTOS, D.; ROCHA, P. A. Estudando o português tal como é usado: o serviço AC/DC. In: PARDO & NUNES.(Eds) The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), 2009. Disponível em <http://nilc.icmc.usp.br/til/stil2009_English/Proceedings/stil/Costa-57572_1.pdf>. Acessado em: 30 out. 2017.

[16] Veja-se o seguinte ponto levantado na lista de discussão do projeto Universal Dependencies: "(...)For example, should "old" be ADJ or NOUN in 'the old are happy'?".

BEAUGRANDE, R. de. Descriptive Linguistics at the Millennium: Corpus Data as Authentic Language. In: *Journal of Language and Linguistics* 1(2), 91-131. 2002. Disponível em: <http://webspace.buckingham.ac.uk/kbernhardt/journal/1_2/beaugrande1_2.html>. Acessado em: 30 out. 2017.

DERRIDA, J. *Gramatologia*. São Paulo: Perspectiva, 2008.

LEECH, G. Corpora and theories of Linguistic performance. In: SVARTVIK, J. (Ed.) *Directions in corpus linguistics: proceedings of Nobel symposium 82*. Berlin e New York: Mouton de Gruyter, pp.125-148. 1992.

MANNING, C., SCHUTZE, H. *Foundations of statistical natural language processing*. Cambridge: MIT press, 1999.

MORETTI, F. Conjectures on world literature. *New Left review* 1, Jan-Feb 2000, p. 54-68.

PAIXÃO de SOUZA, M. C. P. A Filologia Digital em Língua Portuguesa: alguns caminhos. In: GONÇALVES e BANZA, Ana Paula Banza (Eds.). *Património Textual e Humanidades Digitais: da antiga à nova Filologia*. Évora: CIDEHUS, 2013. Disponível em <http://dspace.uevora.pt/rdpc/bitstream/10174/10468/1/e-book.pdf>. Acessado em: 30 out. 2017.

PINTO. J.P. Práticas contra-disciplinares na produção do conhecimento lingüístico. In: MAGALHÃES, José Sueli de; TRAVAGLIA, Luiz Carlos. (Org.). *Múltiplas perspectivas em Lingüística*. 1ª ed. Uberlândia, MG: Edufu, 2008, v. , p. 50-56.

PENNYCOOK, A. Os limites da linguística. In: SILVA & RAJAGOPALAN (Eds). *A linguística que nos faz falhar.* São Paulo: Parábola, 2004 p. 39-43.

RADEMAKER, A., CHALUB, F., REAL, L., FREITAS, C., BICK, E. e de PAIVA, V. Universal Dependencies for Portuguese. *Proceedings of the International Conference on Dependency Linguistics* (2017). No prelo.

SAMPSON, G. Thoughts on Two Decades of Drawing Trees. In: Abeillé (Ed). *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Springer Netherlands, 2003, p.23-41. Disponível em <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.7118&rep=rep1&type=pdf>. Acessado em: 30 out. 2017.

SANTOS, D. Literature studies in Literateca: between digital humanities and corpus linguistics. No prelo. 2017

SANTOS, D. Gramateca: corpus-based grammar of Portuguese. In: Baptista, J. et al. (Eds). *Computational Processing of Portuguese:11th International Conference (PROPOR 2014)*. Germany: Springer, 2014, p. 214-219. Disponível em <http://www.linguateca.pt/Diana/download/gramateca.pdf>. Acessado em: 30 out. 2017.

SANTOS, D. Podemos contar com as contas?. In: ALUÍSIO, S. & Tagnin, S. (Eds.), New *Language Technologies and Linguistic Research: A Two-way Road.* UK: Cambridge Scholars Publishing, 2014, pp. 194-213.

SANTOS, D. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa* 3 (2), 2011, pp. 113-128.

SAUSSURE, F.de *Curso de Linguística Geral*. São Paulo: Cultrix, 2006.

SINCLAIR. J. Naturalness in Language. In: Aarts J. & Meijs, W. (Eds). *Corpus Linguistics.* Amsterdam: Rodopi. p.203-210, 1983.

SINCLAIR, J. 1991. *Corpus, concordance, collocation*: Describing English language. Oxford University Press. 1991.

SPÄRCK-JONES, K. Computational linguistics: what about the linguistics?. *Computational Linguistics*, Volume 33, n. 3, p.437-441. 2007. Disponível em <http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.3.437>. Acessado em: 30 out. 2017.

WITTGENSTEIN, L. *Investigações filosóficas* (1953). São Paulo: Abril Cultural. (Col. Os Pensadores – trad.: José Carlos Bruni), 1984.

Abstract

Linguistic studies and digital Humanities: corpus and decorporification

Since Saussure, linguistic science relies on the assumption that there is a homogeneous language underlying variation. This paper seeks to refute this assumption, using as methodology statistical explorations in large electronic corpora. *Theories are narratives that try to organize the data at our disposal, and it is therefore reasonable that new data (different both in quantity and quality) produce new narratives. From the perspective presented here, we see language being regular and irregular, without center or periphery. Irregularity is inescapable; language is at once complicated and simple.*

Keywords: *Digital humanities. Post-structuralism.* Corpus. *Language instability. Computational linguistics*