CHAPTER ELEVEN

PODEMOS CONTAR COM AS CONTAS?

DIANA SANTOS[*]

1. Preâmbulo

Este texto tem uma intenção didática, embora não para um ensino na sala de aula, de forma a preparar o terreno para construir uma gramática descritiva do português com base em métodos computacionais e apoiada em corpos.

Durante muito tempo desejei que houvesse um texto em que eu pudesse aprender a usar métodos quantitativos com corpos, que me resolvesse as minhas dúvidas e me apontasse a boa prática. Com a idade, apercebi-me de que, exatamente os meus problemas e dúvidas, ninguém tem, e que eu teria de procurar eu mesma e fazer a síntese de que precisava. Este texto é o primeiro resultado deste meu processo de aprendizagem, que naturalmente não estará acabado enquanto eu viver. Mas a simples escrita das primeiras conclusões é inestimável, tanto pelo retorno que espero obter como pela necessidade de escrever um texto convincente.

De forma a ser relativamente fiel à minha participação no ELC em São Carlos, apresentarei como aperitivo (muito brevemente) alguns dos linguistas quantitativos, ou estatísticos linguísticos, que fizeram história na secção 2, e como a sua contribuição continua a ser problematizada – para mostrar que estamos longe de um consenso ou da possibilidade de um conhecimento indiscutível. Depois entro num modo mais descritivo e, ao mesmo tempo que apresento alguns conceitos importantes, dedico-me à aparentemente trivial questão de medir frequências da passiva em português, ilustrando as várias decisões envolvidas, na secção 3.

[*] Universidade de Oslo & Linguateca.

2. Era uma vez… dois homens e uma mulher

A estatística não é uma ciência exata, ou melhor, o uso de métodos estatísticos nas ciências (chamado estatística inferencial ou modelos estatísticos – por oposição à teoria das probabilidades, que é um ramo da matemática) é uma ferramenta que ajuda a medir a qualidade de modelos ou hipóteses, mas que não dá respostas, nunca, por si suficientes. Nada melhor para o mostrar do que relatar um pouco a própria história do seu desenvolvimento e apontar três pessoas importantes na sua génese e disseminação. George Udny Yule (1871-1951) foi um matemático britânico (inicialmente engenheiro) que, após muitos anos de ensino e investigação em estatística, decidiu dedicar-se a resolver problemas da área das letras com o aparato quantitativo que ele conhecia e tinha ajudado a desenvolver, produzindo a obra Yule (1944)[55]. Se tal mudança de vida foi o resultado de dissabores relacionados com diferenças de opinião entre estatísticos (é conhecida a forma como Pearson tratava os seus opositores[56], e Yule foi um dos alunos de Pearson), ou – como a história oficial indica – motivada por razões de saúde, o que é certo é que o seu entusiasmo e a sua dedicação aos problemas a que se dedicou foram uma bênção para a nossa área.

Ao mesmo tempo, mas do outro lado do oceano e sem conhecimento dos trabalhos de Yule, um linguista americano, George Kingsley Zipf (1902-1950), professor de língua alemã em Stanford, dedicava a sua vida profissional a um sonho que contrasta vivamente com a atitude de Yule, com as obras Zipf (1935, 1947)[57], mas que o levou a contagens e a investigações empíricas muito parecidas. Devido à abrangência dos seus

---

[55] Escolhi aqui as frases que me parecem apresentar melhor o seu interesse ao escrever o livro: *This book arose from a desire to study a particular vocabulary in a case of disputed authorship. When I had advanced some way in that particular study, it became only too clear into how thorny a field of statistics I had strayed. (p. ix) […] The vocabulary and diction of Thomas à Kempis are discussed as evidence. These discussions left in my mind a sense of inadequacy. They deal with such details [...]; but they give no faintest notion as to what his vocabulary is really like as a whole.*

[56] Ver por exemplo Agresti (1996).

[57] Eis como Zipf (1949) apresenta os seus objetivos, na sua obra maior, *Human Behavior and the Principle of Least Effort*:

- *Disclosure of some fundamental principles that seem to govern important aspects of our behavior, both as individuals and as members of social groups*
- *Discover the nature of the underlying principles that govern our conduct*

interesses, Zipf tornou-se rapidamente uma referência incontornável em vários meios científicos, desde o processamento de linguagem natural (que surgiu muito mais tarde) até ao planeamento de cidades (Buescu, 2011). Mas, como é comum, com a fama também vem a citação incorreta e a crítica, e são provavelmente poucos aqueles que hoje em dia leem de fio a pavio um livro de 1000 páginas com cerca de uma centena de assuntos díspares medidos rigorosamente por várias dezenas de colaboradores ou alunos seus. Seja como for, o interesse pelo trabalho de Zipf deu origem a um volume especial a ele dedicado pela revista *Glottometrics* em 2002.

O que é particularmente interessante na comparação entre estes dois pioneiros é que, se mostrarmos as descrições dos objetivos dos livros, e dissermos que um foi escrito por um filólogo, e outro por um engenheiro/matemático, toda a gente atribuiria a descrição oposta. O que demonstra, na minha opinião, que as divisões entre letras e ciências ou entre ciências humanas e exatas são muito mais arbitrárias e castradoras do que úteis.

Por outro lado, uma das figuras mais importantes na visualização dos métodos quantitativos, talvez por razões de preconceito em relação ao papel dos sexos, é mundialmente conhecida pelo seu papel humanitário que reabilitou – ou lançou – a profissão de enfermeira, mas as contribuições de Florence Nightingale (1820-1910) para a logística e para a estatística são raramente mencionadas[58].

Não tendo aqui tempo nem espaço para me alongar sobre a história da estatística ou dos métodos estatísticos na linguística, parece-me, contudo, importante salientar que nem Yule nem Zipf foram poupados a críticas pelos seus seguidores (Herdan (1963), sobre Yule, e George Miller, no prefácio da reedição de Zipf (1935), são dois exemplos contundentes), e que um estudo que compara os méritos das suas propostas em relação à medição do vocabulário é Baayen (2001).

Passo agora a descrever e a tentar erradicar dois erros muito frequentes relacionados com métodos quantitativos, antes de apresentar pormenorizadamente alguns conceitos e reflexões. Termino esta parte com uma citação de Guiraud (1960), ainda pertinente passados cinquenta anos:

- *Neither the natural scientist nor the practical social engineer can afford to ignore the power of such preconceptions (...) Nevertheless, to the natural scientist man's preconceptions do not belong to some other world, but instead are further natural phenomena*
- *The expressed purpose of this book [is] to establish The Principle of Least Effort as the primary principle that governs our entire individual and collective behavior of all sorts*

[58] https://en.wikipedia.org/wiki/Florence_Nightingale

La linguistique est la science statistique type; les statisticiens le savent bien; la plupart des linguistes l'ignorent encore. (A linguística é a ciência estatística por excelência, como todos os estatísticos bem sabem… mas a maioria dos linguistas ainda o ignora.)

3. Alguns conceitos importantes, ilustrados pelo estudo da passiva

3.1. Oposição entre qualitativo e quantitativo

Na minha opinião, a dicotomia entre qualitativo e quantitativo é uma falsa questão, porque é preciso atribuir qualidades para se poder contar, ou ter pelo menos uma ideia de magnitude. Além disso, as linguagens naturais misturam de forma linda essa questão de avaliação de quantidade ou qualidade. Vejam que se diz em português *muito lindo* e *muitos carneiros*, e *não vi nada* e *não gosto nada*, etc., etc. Ou seja, o primeiro *muito* está a qualificar/quantificar uma qualidade, a beleza, enquanto o segundo está a quantificar (contar) um conjunto de objetos. E o primeiro *nada* está a contar o que se viu, zero coisas, enquanto o segundo está apenas a qualificar, como forte, uma qualidade (a de que não gosto de uma dada coisa).

De um ponto de vista mais formal, no âmbito da linguística, o artigo de Karlgren (1975) exemplifica soberbamente como essa dicotomia é falsa, mostrando: dados quantitativos para conclusões qualitativas, dados qualitativos para conclusões quantitativas, e como a quantificação (nas hipóteses, nos dados) é complexa e diversificada. Esse artigo devia ser de leitura obrigatória para quem trabalha com língua e computadores.

3.2. Oposição entre métodos linguísticos e métodos estatísticos

Outra falsa dicotomia, ainda mais perniciosa porque amiúde repetida e pelas ilações que dela se podem tirar, é a escolha entre métodos linguísticos e métodos estatísticos no processamento da linguagem natural. Embora nem sequer necessariamente implicada pelos que a invocam, traz por arrasto a interpretação de que para usar métodos estatísticos não é preciso saber linguística, e que para usar métodos linguísticos não é preciso saber estatística. Ambas "conclusões" incorretas e perigosas.

Em primeiro lugar, se estamos a processar a língua, temos de saber linguística. O nosso assunto, o assunto a que estamos a aplicar os nossos métodos de pesquisa, é a língua (e assumo que a linguística é o estudo da língua). Em segundo lugar, para poder aplicar métodos estatísticos, é

preciso: ou ter hipóteses (neste caso, linguísticas), ou ter analisado linguisticamente os dados para os explorar. Quanto mais informação se tem, mais os métodos científicos nos podem ajudar a aumentar o nosso conhecimento. Para variar um pouco o assunto do artigo, e para sublinhar como esta afirmação é válida sempre que esteja em causa a aplicação de métodos estatísticos a qualquer ramo de conhecimento, invoco o exemplo de Van Hoof (2013), em que a autora, interessada no estudo do Império Romano, usa técnicas de análise de redes sociais sobre as cartas de um sábio da época, demonstrando como é preciso ser profundo conhecedor da personagem e da história do período em que viveu para poder aplicar esses mesmos métodos com sucesso.

No sentido inverso, ou seja, a necessidade de ter alguma noção do que contagens em amostras podem implicar, existe a tendência de muitos (para não dizer a maioria dos) linguistas que usam corpos para encarar de forma extremamente simplista as diferenças numéricas (de contagens) como indicadoras de tal ou tal fenómeno, sem nunca sequer imaginar que os números que obtiveram podem não significar rigorosamente nada – porque, por exemplo, a amostra é tão pequena que as variações medidas são simplesmente devidas à sorte.

Outra prática infelizmente comum, já com outro grau de sofisticação, é a aplicação de testes desajustados ao material. Por isso a literatura da linguística quantitativa está cheia de críticas metodológicas e tem relativamente poucas contribuições que vão ao âmago da questão, no sentido de apresentarem métodos desenvolvidos com base nos próprios problemas linguísticos.

Em conclusão, para poder realmente aproveitar os corpos na linguística é preciso na maioria dos casos possuir conhecimento linguístico e conhecimento estatístico; não se pode ficar a meio caminho.

3.3. Repartição (ou distribuição)

Além da questão (naturalmente importantíssima) da exposição à língua autêntica, em grandes quantidades (permitindo, portanto, generalizações e a procura de mais exemplos de um mesmo fenómeno), uma das primeiras propriedades que apreciamos num corpo é a frequência, e a segunda – por muitos considerada tão ou mais importante que a frequência – é a repartição (veja-se, para definições básicas, o meu artigo "Corporizando algumas questões", Santos, 2008, pag. 55).

> *Ou seja, a frequência absoluta de um dado fenómeno é completamente ininterpretável sem relação com o número máximo de casos possíveis*

quando essa contagem foi efetuada. Já a frequência relativa (que é o quociente do número de ocorrências pelo número total) junto com a distribuição por diferentes categorias permitem uma primeira noção sobre a importância e a correlação com estas últimas.

Este é um conceito antigo, usado já pelos primeiros lexicógrafos computacionais para escolher a inserção ou não de uma palavra como vedeta no dicionário (Juilland & Chang Rodríguez, 1964).

No entanto, é muito importante reparar que a noção de repartição ou distribuição envolve certo número de escolhas adicionais: repartição entre que categorias? Categorias extrínsecas ou intrínsecas ao texto, tal como local de origem ou género literário, ou, pelo contrário, princípio ou fim de um texto ou de uma frase?

Tal como é possível fazer malabarismo eleitoral dividindo criteriosamente as freguesias de voto e a forma de atribuir representantes num órgão de soberania[59], também existem – como em todos os modelos quantitativos – várias possibilidades distintas de modelar os dados, que permitem, portanto, detetar, ou não, a influência de fatores distintos.

Neste artigo vou descrever essa questão recorrendo a exemplos retirados da gramática do português, começando por esclarecer alguns conceitos que no meu entender têm sido (ainda) pouco problematizados.

3.4. Frequência (relativa) e a questão das unidades

O primeiro – e intuitivo – conceito associado à quantificação é o de frequência relativa. Para se poder comparar dois números, não há ninguém que duvide de que é preciso conhecer o tamanho do material: se se encontraram dois casos de X em cem palavras, e noutro estudo dois casos do mesmo X em 300 milhões de palavras, embora a frequência absoluta (dois) seja a mesma, é óbvio que a diferença na frequência relativa é abissal.

Contudo, se para este exemplo ilustrativo podemos usar o conceito de palavra como um indicador de tamanho, em muitos casos é preciso uma reflexão mais aturada sobre qual a unidade por que normalizamos, sobretudo se estivermos interessados em fenómenos que por si próprios incluem mais de uma unidade. (Para um tratamento pormenorizado do problema das unidades em geral, aconselho vivamente o texto de Krippendorff (2004)).

---

[59] Veja-se por exemplo http://fr.wikipedia.org/wiki/D%C3%A9coupage_%C3%A9 lectoral

Autores há, aliás, que preferem dividir os textos em bocados/unidades arbitrárias de igual tamanho (por exemplo, em palavras), e fazer as suas contagens sobre essas divisões. Frumkina (1962), citada em Köhler (2012), e Biber (1985) são exemplos dessa metodologia. Contudo, podem existir vários problemas com estes blocos, que em si são definidos em termos de outras unidades.

Um exemplo óbvio, ainda recorrendo à noção de palavra gráfica, é a contagem de expressões com mais de uma palavra (EVP). Se a apresentarmos como o número de EVP por palavra (ou por mil palavras, por exemplo), estamos a recorrer a uma medida notoriamente difícil de interpretar, visto que as contagens não se referem aos mesmos objetos, e, portanto, essa medida não é estritamente uma proporção (embora o pareça ser). Para ser uma proporção, as contagens no numerador e no denominador têm de se referir à mesma unidade. Neste caso, uma medida mais natural seria o número de palavras pertencentes a EVP comparado com o número de palavras total, ou, para poder contabilizar também o tamanho das EVP, um conjunto de medidas para cada tamanho de EVP encontrado: a proporção de palavras pertencentes a EVP de tamanho 2, de tamanho 3, etc.

Para concretizar, veja-se a figura 1 com um texto muito pequeno e diversas contagens sobre ele.

RIO - Uma edição especial do tradicional jogo de tabuleiro Banco Imobiliário, carregada de elogios a obras e programas do prefeito da cidade, Eduardo Paes (PMDB), está sendo distribuída em escolas públicas municipais do Rio de Janeiro.
O jogador não compra mais imóveis em bairros tradicionais de São Paulo ou do Rio, como na versão tradicional, mas passa a investir seus recursos em iniciativas como BRTs (via exclusiva para ônibus), Clínica da Família, Museu do Amanhã, Bairro Carioca, entre outras da administração Paes.

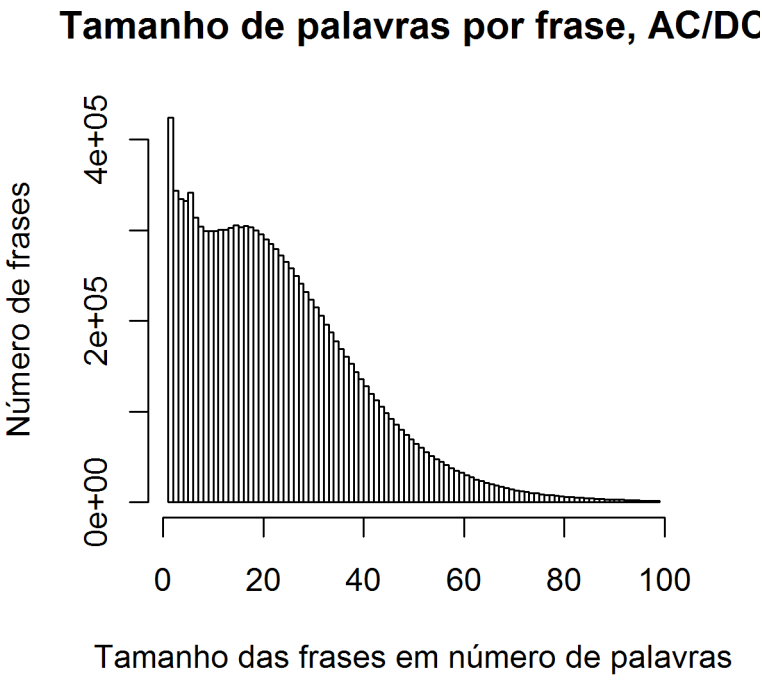
Número de palavras: 82 (pontuação não foi contada)
Número de EVP (sublinhadas): 7
Número de palavras pertencentes a EVP:17
Número de palavras pertencentes a EVP de tamanho dois: 8
Número de palavras pertencentes a EVP de tamanho três: 9

Fig. 1: Excerto de uma notícia breve publicada pelo jornal Estadão, gentilmente cedida pela equipa do CSTNews

Outro exemplo, mais comum, mas que incorpora um problema semelhante, diz respeito à frequência de uma construção sintática ou semântica. A pergunta, aqui, para chegar a proporções, é a mesma: o que

se coloca no denominador? Ou: qual é o termo de comparação sobre o qual se conta? É possível propor proporções defensáveis?

Imaginemos que estamos interessados na frequência da passiva em português, ou na frequência do uso do futuro (simples). É interessante salientar que o exemplo da passiva é bastante frequente em textos de estatística na língua, veja-se Baroni & Evert (2008), Halliday (1991) e Köhler (2012), além de ser um dos muitos fatores em Biber (2005).


Tamanho de palavras por frase, AC/DC
Número de frases
0e+00
2e+05
4e+05
0
20
40
60
80
100
Tamanho das frases em número de palavras

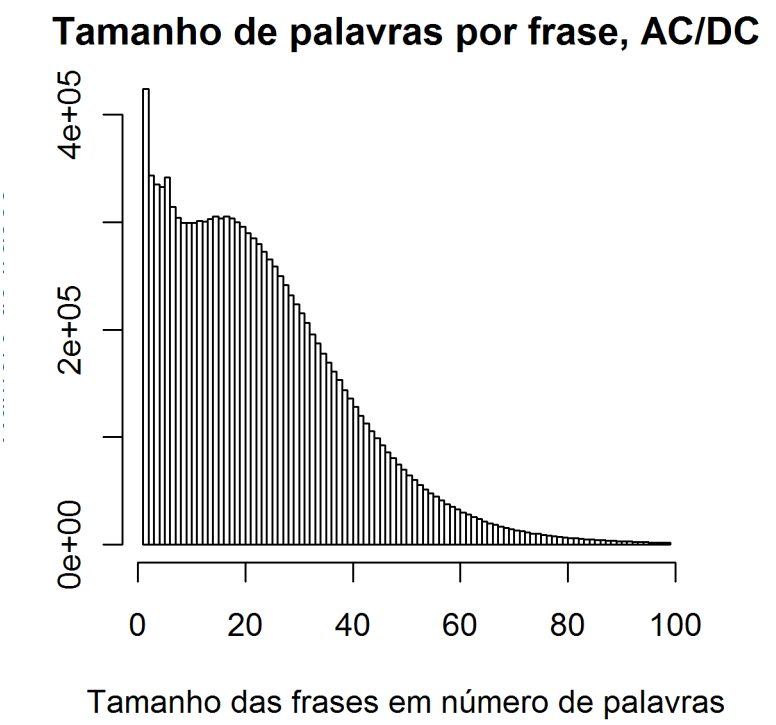


Figura 2a: Distribuição do tamanho das frases de menos de 100 palavras no AC/DC, por número de palavras

Frequentemente usa-se o número de frases como quantificador do universo, porque é relativamente fácil contá-las.[60] Mas, dado que em português o número de orações numa frase é geralmente muito maior do que um (ver figura 2a para a distribuição de frases por número de palavras, e figura 2b para a a distribuição de frases por número de orações no

[60] Fácil, e fácil… a questão da contagem das palavras não é certamente consensual, visto que a atomização implica muitas pequenas decisões, e em português até já medimos a diferença entre vários sistemas computacionais nas Morfolimpíadas (Santos et al. 2003). Além disso, existe o problema das contrações e dos clíticos (mesoclíticos e enclíticos) e das locuções e nomes próprios, que levou a uma atomização do AC/DC diferente da do PALAVRAS (Santos & Bick, 2000), sem contar com as siglas e os números: 306, *trezentos e seis*, quantas palavras são?

AC/DC[61]), não só a "proporção" é incorreta como poderia facilmente dar azo a números maiores do que um. (No caso de um texto todo na passiva, com 2,4 orações por frase, teríamos 2,4 passivas por frase, o que não é naturalmente uma proporção.)

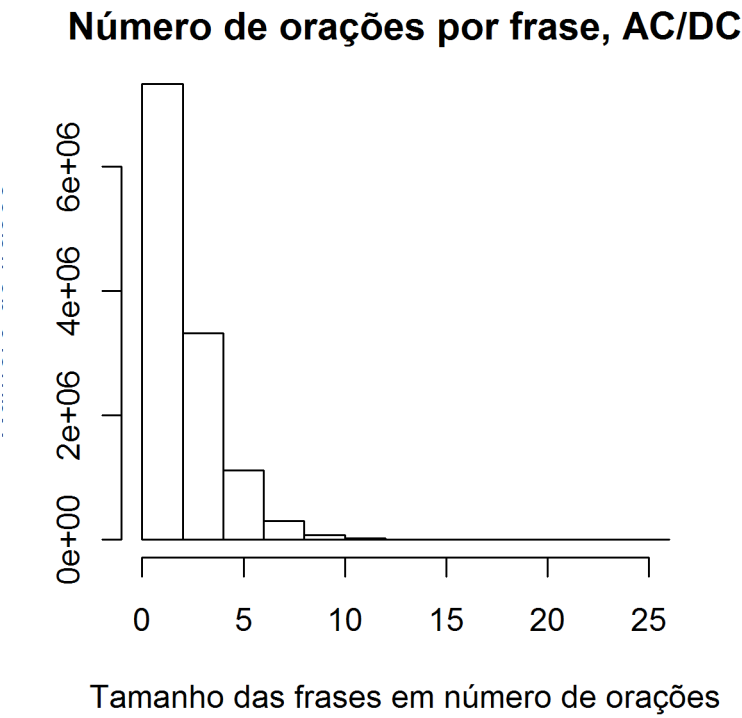


Figura 2b: Distribuição do tamanho das frases de menos de 100 palavras no AC/DC, por número de orações

Visto que qualquer oração pode ser classificada como passiva ou não, podíamos, contudo, propor como uma boa medida para o grau de passiva a proporção de orações na passiva.

Repare-se, antes disso, como seria enganadora uma medida de passivas por palavras. Em primeiro lugar, porque sendo a passiva em português constituída por um auxiliar e o verbo principal no particípio passado, a

[61] http://www.linguateca.pt/ACDC/

própria existência da passiva causa na maioria dos casos o aumento de número de palavras.

Em segundo lugar, porque o número de palavras do texto pode não estar relacionado com o número de orações. Senão vejamos: Num texto com 20 orações de 20 palavras cada, metade na passiva, e noutro texto com 100 orações com quatro palavras cada, metade na passiva, a nossa medida por oração considerá-los-ia semelhantes (grau de passiva: 50%), enquanto que, se a medida fosse passiva por número de palavras, teríamos 10/400 para o primeiro texto, e 50/400 para o segundo!

Mas, voltando à nossa medida da passiva como proporção em orações, ainda se poderia ir mais longe e argumentar que não se deveria medir as passivas por oração porque nem todas as orações que não estão na passiva o poderiam estar: como se sabe, há verbos em português que não podem ser passivizados, como *gostar*, *ser*, *desmaiar*, etc. Há muitas razões distintas para esta impossibilidade, sobre as quais não vou agora espraiar-me, mas um linguista mais consciencioso poderia defender que levássemos em conta o verbo principal da oração em questão.

Ou seja, poderíamos restringir o nosso índice de passiva de forma a apenas cobrir os casos dos verbos principais que poderiam ser passivados (e, da mesma forma, "ativados", nos casos, muito mais raros, de um verbo apenas poder ser usado na passiva).[62]

Esta sugestão tem, contudo, um reverso, que me parece importante salientar: de facto, um índice quantitativo depende do objetivo para o qual foi desenhado. Se o quiséssemos para comparar dois autores, poder-se-ia argumentar que a própria escolha lexical de um autor por verbos mais (ou menos) passiváveis contribui decisivamente para o seu estilo, e que portanto se poderia falar de uma tendência para a passiva – ou para a ativa – de um autor, independentemente do nível mais fino dos itens lexicais escolhidos.

Por outro lado, seguindo um raciocínio semelhante, podemos estar interessados na frequência da passiva, não em geral, mas por tipo de oração. Por exemplo, uma das maneiras como eu apresento a passiva aos meus alunos de português é mostrando como as orações relativas ficam

[62] Não pretendo, naturalmente, entrar aqui nos meandros do que é considerado passiva ou não – é um assunto bastante discutido, e polémico, na gramática do português (Casteleiro, 1981, Ranchhod, 1990, Peres & Móia, 1995, Barreiro, 1998, Afonso, 2008, etc.). Estou a assumir que as opções linguísticas do PALAVRAS (Bick, 2000) estão corretas, e que o grau de correção do sistema automático é suficientemente elevado para o erro ser negligenciado. Alternativamente, poderíamos fazer estas contagens em material revisto por linguistas, como a Floresta (Bick et al., 2007).

mais fáceis de compreender. Assim mostro-lhes os seguintes exemplos, em que o asterisco significa pouco natural, e em que a segunda possibilidade é descrita como mais idiomática:

** O presidente subiu ao poder. O povo elegeu o presidente.*
1. O presidente que o povo elegeu subiu ao poder
2. O presidente que foi eleito pelo povo subiu ao poder

** O homem entrou na casa. O pai dele comprou a casa dois anos antes.*
1. *O homem entrou na casa que o pai dele tinha comprado dois anos antes*
2. *O homem entrou na casa que tinha sido comprada dois anos antes pelo pai*

A pergunta fica: isto é apenas um bom argumento pedagógico, ou na realidade a frequência da passiva é maior em orações relativas? Para poder responder a esta pergunta com base em corpos, temos primeiro de pensar como medir a diferença de frequências.

3.5. Comparação de duas frequências

Em primeiro lugar, o que é que se compara? O índice da passiva só em orações relativas, com o índice geral (incluindo todas as orações)? Ou comparando com o resto das orações?

E já agora, as orações participiais: são passivas, ou ainda outra coisa?

E as orações sem verbo? Para que lado contam?

Parece-me mais natural retirar todos os casos complicados da comparação, e medir a proporção de passiva em orações relativas, e a proporção de passiva noutras orações que tenham verbo, mas excluindo também as orações participiais (que podem ser consideradas como uma passiva "despida").

Mas, mais importante do que aceitar ou não estes argumentos e forma de proceder, é reparar como uma (aparentemente) "simples" hipótese implica tantas decisões ~~em~~ ao nível linguístico antes de se poder proceder à contagem e depois eventualmente aplicar métodos estatísticos.

Na f~~F~~igura 3, alguns números são apresentados, com base no conteúdos corpos do AC/DC em junho de 2013. Para uma descrição deste serviço e dos corpos a que dá acesso, veja-se Santos (2011) e Santos (201~~3~~4), enquanto em Santos (2012) algumas formas iniciais de os explorar são aventadas.

Ocorrências Verbos diferentes

Verbos principais:	30.049.613	41.975
Passivas:	1.905.338 (6,3%)	7.584 (18,1%)
Particípios passados (outros):	3.526.785	
Verbos com objeto direto:	13.191.083	
Verbos com OD ou passiva:	15.096.421 (12,6%)	
Orações relativas:	4.949.031	
Orações relativas passivas	270.564 (5,5%)	
Verbos T em orações relativas:	1.451.078	
Orações relativas T ou passivas:	1.721.642 (15,7%)	

Figura 3: Contagens de diferentes subconjuntos nos corpos do AC/DC (as percentagens são sempre das passivas em relação ao grupo anterior)

Da figura 3 constatamos que, dependendo da forma como operacionalizamos o problema, podemos concluir que há menos (5,5% comparado com 6,3%) ou mais (15,7% comparado com 12,6%) passivas nas orações relativas. Ou podemos até concluir que não podemos concluir nada, visto que a própria contagem de verbos transitivos (T) depende das ocorrências no corpo e não da sua própria "personalidade", e muitos verbos transitivos podem aparecer sem o objeto expresso.

E, de qualquer maneira, convém relembrar que essas classificações são feitas automaticamente e que por isso os números não são exatos. Em alguns casos o PALAVRAS – ou o programa do AC/DC que marca as passivas – sobre-analisa (primeiro bloco), noutros (segundo bloco) sub-analisa:

o que eu não tinha feito era posto empenho suficiente
E estão derrubado quilômetros e quilômetros (Erro no próprio texto)
Minha mãe era advogada voluntária
Francisco Rezek não é um exagerado otimista
Em causa está alegada falta de diálogo (Título)
o peruano Baroni , era presa fácil
que fosse assim – e o vestido preto de meia

Será criado um organismo
está aliado também às fugidas noturnas
susceptíveis de serem convertidas sistematicamente
mas nós não estávamos interessados
descobrir onde estavam detidos
essa possibilidade foi a escolhida pelo Brasil

3.6. A questão da possibilidade de repetição

Antes que estas manipulações pareçam simples truques de prestidigitação, lembremos que a análise e respetivas contagens que apresentámos aqui pode ser repetida e refinada por todos quantos quiserem estudar o problema, reusando os nossos dados.[63] O que, aliás, é um pressuposto do método científico: a partilha de dados.

Mas, claro, esta opção só é possível se tivermos acesso ao material. Se estivermos a comparar com resultados na literatura, teremos de tentar escolher as mesmas opções, ou tentar modelar, a partir dos nossos dados, qual a influência de opções diferentes.

Imaginem, assim, que tínhamos dados de outro estudo, que indicavam que a frequência das passivas, medida em passivas por frase, era de 0,30.

Teríamos de, primeiro, estimar o número de orações por frase no nosso material, e aplicar esse valor ao número citado, ficando, por exemplo, com 0,15 passivas por oração. Depois, estimaríamos o número de orações sem verbo, se suspeitássemos que tal correção não tinha sido feita, corrigindo talvez para 0,18, etc., etc. Seja como for, haveria sempre muitos casos que não poderíamos confirmar se eram diferentes ou simplesmente tinham sido analisados de maneira diferente.

Por isso, o ideal é ter acesso ao material sobre o qual as contagens do primeiro autor tivessem sido feitas, para se poder tornar a contar com outra metodologia se necessário fosse (cf. Santos & Oksefjell, 1999). A questão de poder repetir as contagens é essencial e nunca é possível fazê-lo sem ter acesso aos próprios dados.

3.7. Interação entre dois fenómenos linguísticos

Vejamos agora se quiséssemos estudar a interação do tempo verbal com a passiva. Ou seja, por enquanto de forma pouco rigorosa ~~por enquanto~~: Será que a passiva tem tempos preferidos ou tempos preteridos? Ou será que um dado tempo verbal tem preferência pela passiva? Ou ainda, será que há interação com outros aspetos, por exemplo, a progressiva?

Começo desde já por alertar para o seguinte: a terceira pergunta é diferente em género das duas primeiras, que continuam apenas a pedir proporções:

[63] Em anexo, indicamos os comandos exatos assim como os testes aplicados.

A primeira refere-se à comparação de proporções de tempos verbais em orações ativas e passivas, e a segunda à comparação da proporção de orações na voz ativa e passiva para cada tempo verbal.

A terceira, por outro lado, pergunta simplesmente se os dois fenómenos são independentes ou relacionados. Isto porque, para cada oração se pode contabilizar separadamente se está na passiva e se está na progressiva, obtendo-se a tabela 1, que indica que a interação entre os dois fenómenos é significativa.

Tabela 1: Tabela de contingência entre passiva e progressiva, repetida 3 vezes. A cinzento, estão as percentagens por coluna (referentes à progressiva); mais à direita as percentagens por linha (referentes à passiva)

	pProgressiva	não				
passiva	40.284	1.865.054	12%	0,6%	2,1%	97,9%
não	305.229	295.962.353	87%	99,4%	0,1%	99,9%

Mas voltemos por agora às duas primeiras perguntas. Em ambas, as grandezas que nos interessam são frequências de ocorrência de uma oração – tanto na passiva como num dado tempo. Contudo, é preciso notar que cada pergunta se refere a proporções diferentes, que é importante não confundir. Por outras palavras: em termos estatísticos, temos uma pergunta no universo dos tempos, e uma pergunta no universo da passiva, e esses universos são diferentes. Por isso, as nossas contas e as nossas possíveis conclusões são diferentes. Nas figuras 4 e 5, mais uma vez alguns números do AC/DC são apresentados:

	Total	Passiva	%
Presente do indicativo	12.067.989	542.305	4,5
Imperfeito do indicativo	2.090.276	111.598	5,3
Perfeito do indicativo	6.684.647	565.087	8,4
Pretérito perfeito composto do ind.	170.737	16.154	9,5
Mais que perfeito	1.691.666	196.896	11,6
Presente do conjuntivo	1.050.565	56.750	5,4

Fig. 4: percentagem de casos na passiva, por tempo

Da figura 4, podemos concluir que há tempos com mais preferência para a passiva, e que as diferenças são significativas. E da figura 5 resulta claro que tempos diferentes dão diferentes quinhões à passiva.

	Total	%
Perfeito	565.087	29,7
Presente do indicativo	542.305	28,5
Infinitivo	354.355	18,8
Futuro do indicativo	167.275	8,8
Imperfeito	111.598	5,8
Presente do conjuntivo	56.750	3,0

Fig. 5: percentagem dos tempos, na passiva (1.905.338 casos)

Mas mais uma vez é preciso notar a miríade de opções necessárias, desta feita para contar tempos (verbais). Por exemplo, porque o pretérito perfeito composto (PPC) é um tempo especial, separei-o. E porque o mais que perfeito (MQP) tem uma forma sintética e outra analítica, amalgamei ambas. Estas escolhas, embora na minha opinião perfeitamente defensáveis, diminuem significativamente os valores do presente e do imperfeito (quando o verbo é *ter*).

Além disso, repare-se que existem vários tempos compostos que ainda não têm um nome (alguns não têm um nome consensual, outros não têm sequer nome), como é o caso do futuro com *ir*, ou de vários aspetualizadores (veja-se Freed 1979, Santos 1995), em que o caso do *ir* é outra vez um dos mais notórios.

Para esclarecer o que pretendo indicar com o parágrafo anterior, abro aqui um parêntesis relativo ao verbo *ir*: existe uma riqueza extraordinária de sentidos associada ao uso de *ir* (e de *estar*) como auxiliar em português, como as seguintes frases (inventadas por mim) demonstram.

Eu ia caindo
Eu ia a sair, quando me lembrei de que não lhe tinha telefonado.
Eu fui comprar o livro e por isso não o encontrei
Eu ia comprar o livro, mas estava esgotado.
Ele lá ia fazendo o que lhe pediam.
As pessoas iam chegando, e viam aquele espetáculo.
As pessoas foram chegando, até que às duas já havia quorum para a reunião.
As pessoas vão chegando, e procuram o seu lugar.
Vai-se andando…

Estou convencida de que não é, pois, ainda possível, sem um estudo mais aturado destes casos, distinguir entre gradualidade (forma progressiva), intenção, quase realização, ou simplesmente duratividade – sem esquecer que, por vezes, mais do que um desses valores pode e deve ser atribuído, dada a vagueza essencial da língua (Santos, 2007).

Como lidar com este problema? Uma hipótese seria retirar todos os casos do auxiliar *ir* das contagens dos tempos (e, portanto, também das passivas). Outra seria atribuir tempos / sentidos diferenciados aos casos do *ir*, e outra ainda seria contabilizá-los nos tempos do primeiro auxiliar (*iam chegando* seria imperfeito, *fui comprar* seria perfeito). Esta será uma decisão que terá de ser tomada quando realmente desenharmos a gramática baseada em corpos. Fecho aqui o parêntesis, confessando que os números apresentados não separam o verbo *ir*, exceto quando se refere ao futuro.

Outra decisão difícil, seja como for, é a contagem do número de orações quando há aspetualizadores, visto que uma oração nova implica uma possibilidade de passivização nova.

No entanto, os próprios aspetualizadores não podem estar na passiva e ao mesmo tempo continuarem a funcionar como aspetualizadores,[64] por isso podemos considerar que estamos em presença de uma oração só:

Ele está a começar a ser muito apreciado pelos colegas.
Ele deixou de ficar magoado com os comentários dela.
O muro acabou de ser pintado ontem.

Convém de qualquer forma salientar que é perfeitamente possível, e idiomático, ter estruturas (superficialmente) análogas às anteriores com dupla passiva, e que são analisadas da mesma forma pelo PALAVRAS:

Ele ter sido espancado não foi nada apreciado.
Ele não foi ensinado a ser fotografado.
Ele não estava habituado a ser ludibriado.
Ele foi proibido de ser entrevistado.
Ele estava cansado de ser interrompido.

Estes exemplos foram escolhidos com uma segunda intenção: a de mostrar que a própria qualificação de passiva, ou melhor, da escolha de atribuição da classificação "passiva" exige várias decisões que são geralmente chamadas do foro qualitativo. E que, portanto, é absolutamente impossível, e impensável, fazer estudos de gramática baseados em corpos sem fazer decisões linguisticamente motivadas e bastante complexas, que por sua vez nos poderão permitir descobrir propriedades da língua sobre as quais ainda não tínhamos refletido.

[64] Esta é uma afirmação que faço baseada na minha competência linguística, mas que será preciso confirmar.

3.8. Consequências para a linguística com corpos

Pode acontecer que, a braços com todas estas decisões, no fim de contas os números acabem por não importar, comparados com as várias sistematizações conseguidas, mas certamente que a necessidade de olhar para os dados e para as alternativas possíveis é extremamente enriquecedora em termos linguísticos – e, se as análises e a anotação forem tornadas públicas, podem servir de base para futuros estudos.

Este desbravar das contagens e das qualificações/categorizações necessárias é um pré-requisito para trabalho de qualidade numa gramática baseada em corpos. Como já defendido em Santos (2012), se essa gramática for baseada em dados públicos, será um progresso considerável em relação à sua predecessora para o inglês (Biber et al., 1999), na qual naturalmente me inspirei[65] mas que pretendo, em equipa, ultrapassar sempre que tal for possível. Os autores da gramática inglesa, quinze anos antes, tinham certamente ao seu dispor menos poder computacional e técnicas estatísticas menos sofisticadas, além de não existir, para o inglês, um analisador sintático da craveira do PALAVRAS.

Resta-me, pois, concluir: é minha convicção profunda que, para contar, é preciso categorizar, e ao categorizar descobrimos qualidades da gramática de uma língua com que nunca teríamos deparado sem esse esforço de sistematização. Por isso, os métodos quantitativos e qualitativos são duas faces da mesma moeda para a compreensão de como a língua, a ferramenta mais complexa de que a humanidade dispõe, funciona.

Referências

Afonso, Susana. 2008. *The Family of Impersonal Constructions in European Portuguese. An Onomasiological Constructional Approach*. Tese de doutoramento, University of Manchester.

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*, John Wiley and Sons.

Baayen, R. Harald. 2001. *Word frequency distributions*. Kluwer Academic Pusblishers.

[65] Outra fonte a que recorreremos constantemente e que, portanto, será uma peça fundamental na descrição do português é a série de volumes do NURC sobre a norma urbana culta nas cidades brasileiras, e que é a primeira gramática do português (ou manancial de estudos gramaticais) baseada em corpos, veja-se Varejão (2009) para uma perspectiva histórica.

Baroni, Marco ~~and~~ & Stefan Evert.2008. “Statistical methods for corpus exploitation”. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, artigo 36. Berlim, Mouton de Gruyter.

Barreiro, Anabela Marques. 1998. "Propriedades Sintáctico-Semânticas dos Particípios Passados em Português Europeu", Tese de Mestrado, Universidade Nova de Lisboa.

Biber, Douglas. 1985. "Investigating macroscopic textual variation through multifeature/multidimensional analyses". *Linguistics* 23, 2, pp. 337-360.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & E. Finegan. 1999. *The Longman grammar of spoken and written English*. 1999, ~~London:~~ Longman.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Bick, Eckhard, Diana Santos, Susana Afonso & Raquel Marchi. 2007. "Floresta Sintáctica: Realidade ou ficção?", in Santos, Diana (org.) *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, IST Press, 2007, pp. 291-300.

Buescu, Jorge. 2011. “Matemática, a cidade e a vida”, capítulo 14 de *Casamentos e outros Desencontros*, Gradiva, pp.139-46.

Casteleiro, João Malaca. 1981. *Sintaxe transformacional do adjectivo: regência das construções completivas*. INIC, Lisboa.

Freed, Alice F. 1979. *The Semantics of English Aspectual Complementation*, Dordrecht, D. Reidel.

Frumkina, Revekka Markovna. 1962. “O zakonachraspredelenija slov I klassov slov.” In Mološnaja, Tatj’ana N. (ed.), *Strukturno-tipologičeskie issledovanija*. Moscovo, Academia da URSS, pp. 124-133.

Glottometrics: to honor G. K. Zipf, 3, 2002, RAM-Verlag, http://www.arteuna.com/talleres/lab/ediciones/libreria/Glottometrics-zipf.pdf

Guiraud, Pierre. 1960. *Problèmes et Méthodes de la statistique linguistique*, Paris, P.U.F.

Halliday, M.A.K. 1991. "Corpus studies and probabilistic grammar" in Aijmer, Karin & Bengt Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, pp.30-43.

Herdan, Gustav. 1963. "A method for the quantitative analysis of language mixture», *SMIL* 2, 1963, pp. 110-123.

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency Dictionary of Spanish Words* (The Romance Languages and their Structures, First Series, S 1). The Hague: Mouton & Co.

Karlgren, Hans. 1975. “Quantitative models – of what?”, *Statistical Methods in Linguistics, SMIL* 1975, pp. 25-31.

Katz, Slava M. 1996. "Distribution of content words and phrases in text and language modelling", *Natural Language Engineering* 2 (1996), pp. 15-59.

Köhler, Reinhard. 2012. *Quantitative syntax analysis*. De Gruyter.

Krippendorff, Klaus. 2004. *Content Analysis: an introduction to its Methodology*. Sage Publications, 2ª edição. 1ª edição: 1980.

Peres, João Andrade & Telmo Móia. 1995. *Áreas Críticas da Língua Portuguesa*. Lisboa, Caminho.

Ranchhod, Elisabete Marques. 1990. *Sintaxe dos predicados nominais com Estar*, INIC, Lisboa.

Santos, Diana. 1995. "On grammatical translationese", in *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics* Helsinki, 29-30th May 1995, compiled by Kimmo Koskenniemi, pp. 59-66.

—. 2007. "O modelo semântico usado no Primeiro HAREM". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, pp. 43-57.

—. 2008. "Corporizando algumas questões". In Stella E. O. Tagnin & Oto Araújo Vale (orgs.), *Avanços da Lingüística de Corpus no Brasil*, Editora Humanitas/FFLCH/USP, São Paulo, pp. 41-66.

—. 2011. "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties". in J.B. Johannessen (ed.), *Language Variation Infrastructure. OSLa: Oslo Studies in Language* 3.2 (2011), pp. 113-128.

—. 2012. "The next step for the translation network". In Diana Santos, Krister Lindén & Wanjiku Nganga (eds.), *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th BirthdayCarlson*. Springer, 2012, pp. 49-62.

—. 20134. "Corpora at Linguateca: Vision and roads taken", Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*, Bloomsbury, no prelo.

Santos, Diana & Eckhard Bick. 2000. "Providing Internet access to Portuguese corpora: the AC/DC project", in Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), pp. 205-210.

Santos, Diana & Signe Oksefjell. 1999. "Using a Parallel Corpus to Validate Independent Claims", *Languages in contrast* 2 (1), pp.117-132.

Santos, Diana, Luís Costa & Paulo Rocha. 2003. "Cooperatively evaluating Portuguese morphology", in Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings*, Springer Verlag, pp. 259-66.

Van Hoov, Lieve. 2013. “SNA & ancient literature: Libanius' Epistolary Ego-Network”, http://de.digitalclassicist.org/berlin/files/slides/dcsb_van-hoof_22012013.pdf

Yule, George Udny. 1944. *The statistical study of literary vocabulary*, Cambridge University Press, 1944.

Varejão, Filomena de Oliveira Azevedo. 2009. "O português do Brasil: Revisitando a História", *Cadernos de Letras da UFF – Dossiê: Difusão da língua portuguesa* 39, pp. 119-137.

Zipf, George Kingsley. 1935. *The Psycho-Biology of Language.* Cambridge Mass., 1935. Reissued in 1965 with a preface by George Miller.

—. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley Press, Cambridge, Mass., 1949.

Anexo: Comandos usados para obter os dados das tabelas

Número de verbos principais: [func=".*MV.*"];
Verbos na passiva: [func=".*MV.*" & temcagr=".*PASSIVA.*"]
Verbos transitivos: (@[pos="V.*"] [pos!="V.*"]*
[func="<ACC.*"])|([func="ACC>"] [pos="ADV"]*
@[pos="V.*"])|(@[pos="V.*" & func=".*ACC.*"]) within s;
Orações relativas: [pos=".*rel"] [func!=".*MV.*"]* [func=".*MV.*"]
within s;
Orações relativas na passiva: [pos=".*rel"] [func!=".*MV.*"]*
[func=".*MV.*" & temcagr=".*PASSIVA.*"];
Orações relativas com objeto direto: ([pos=".*rel" & lema!="como"]
[func!=".*MV.*"]* [func="ACC>"] @[func!=".*MV.*"]*
@[func=".*MV.*"])|([pos=".*rel" & lema!="como"] [func!
=".*MV.*"]* @[func=".*MV.*"] [pos!="V.*"]*
[func="<ACC.*"])|([pos=".*rel"] [func!=".*MV.*"]* [func="ACC>"]
[func!=".*MV.*"]* @[func=".*MV.*" & temcagr=".*ACC.*"]) within
s;
Orações com PPC: [temcagr=".*PPC.*"]
Orações com PPC na passiva: [temcagr=".*PPC.*" &
temcagr=".*PASSIVA.*"];
Orações com PPC na passiva e na progressiva: [temcagr=".*PPC.*" &
temcagr=".*PASSIVA.*" & temcagr=".*PROG.*"];