COMPARANDO CORPOS ORAIS (TRANSCRITOS) E ESCRITOS USANDO A GRAMATECA

DIANA SANTOS

d.s.m.santos@ilos.uio.no

Abstract

Is there a sensible way to compare oral and written Portuguese using the large number of corpora available in AC/DC, underlying corpus-based Gramateca? In this paper I present what is currently possible with the tools and data available, after discussing whether the comparison makes sense.

I start by providing a bird's eye of the corpus material, showing the huge diversity between the oral corpora. I then pose several questions, namely: what is the amount of second person, its contrast with first person; what kinds of first person plural there are; how to identify the passive, identify emotions, and understand the use of the word então (“*so”*, or “*then”*). *Rather than provide definite answers, my concern is to show how to conduct this kind of studies, pointing out various problems that have to be taken in consideration.*

1. É possível comparar?

Este artigo começou com uma interrogação: faz sentido comparar como falamos e como escrevemos? E, se sim, comparar em relação a quê? Existe uma separação concetual entre as duas linguagens (português oral e escrito)? Ou, simplesmente, algo é mais oral ou mais escrito? visto que, teoricamente, se pode ler alto qualquer texto, e escrever tudo quanto se diz. Até que ponto não há diferença entre o oral e o escrito no sentido de que é o género textual que é o fator, e não o meio de pronunciar/escrever?

Biber (1998) parece-me ter chegado à conclusão de que é o género que provoca as diferenças (pelo menos, as diferenças que ele quis medir), e não a dicotomia oral-escrito, por isso não parece desfasada a ideia de tentar verificar, com base numa classificação por géneros, se os orais e os escritos se encontram próximos. Mas o género, infelizmente, não é algo fácil de

distinguir e delimitar... Por um lado, alguns géneros são orais por excelência, e outros são, ao invés, pensados para serem escritos, mas muitos deles podem ser concetualizados como tendo duas faces como Janus, estando indissociavelmente associados a ambos os meios. Estou a pensar no caso dos *Diários da República* ou das sebentas universitárias, cujo objetivo é – ou era – guardar as palavras faladas dos legisladores ou dos professores; e no caso das peças de teatro, cujo objetivo é serem faladas (ditas, representadas), mas que geralmente são escritas pelo dramaturgo primeiro. E o mesmo acontece no caso dos discursos não improvisados (lidos).

Por outro lado, consoante um texto é falado (por meio de ondas sonoras) ou escrito (através de um código visual) existem propriedades diferentes que podem ser comparadas. Por exemplo, teóricos há que definem o conceito de "palavra fonológica" contraposto à "palavra gráfica". Numa formulação propositadamente ateórica, pode dizer-se que características como ritmo, intensidade e entoação são muito mais difíceis de estudar na língua escrita[1]. Veja-se, contudo, o recente artigo de Galves et al. (2013), que estuda o ritmo em texto jornalístico escrito. Por outro lado, quantidades como número de relativas ou cadeias anafóricas são mais fáceis de identificar em texto escrito. Biber & Gray (2010), comparando o inglês oral com o escrito, vem no entanto desmentir algumas das crenças sobre a complexidade da língua escrita. É relevante fazer trabalhos no mesmo espírito para a língua portuguesa, visto que é sabido que esta, pelo menos na escrita académica, se comporta de maneira muito diferente da língua inglesa (Bennett, 2010).

Usamos, neste trabalho, a Gramateca (Santos, 2014) para tentar uma primeira resposta a algumas destas perguntas, por isso convém começarmos por perguntar se a Gramateca é o ambiente ideal. O AC/DC (Santos & Bick, 2000; Santos, 2011), sobre o qual a Gramateca se apoia, é, sem dúvida a infra-estrutura linguístico-computacional que dá hoje acesso a um maior número de textos anotados em língua portuguesa, graças à confiança da comunidade nos nossos serviços e nas nossas intenções de contribuir para o processamento da língua comum, ao invés de daí tirar partido para ganhos económicos ou políticos. Em julho de 2014,[2] a Linguateca dá acesso a vinte

1 Exceto na poesia, também discutível se constitui um género escrito ou oral: é escrita e concebida para ser lida, ou para ser declamada, ou para ambas?

2 Todos os valores quantitativos deste artigo referem-se a essa data.

e cinco corpos diferentes, criados em tempos diferentes por investigadores diferentes com objetivos diferentes, mas todos aderindo a um formato de interrogação comum e tendo sido enriquecidos com bastante informação sintáctica e semântica adicional: análise sintática do PALAVRAS (Bick, 2000) e análise semântica no que respeita a cores, corpo e emoções.[3]

No entanto, estamos longe de afirmar ser a Gramateca o ambiente ideal para comparar o oral e o escrito em português. Por várias razões: Em primeiro lugar, temos muito mais material escrito, e de forma alguma consideramos que o conjunto dos textos orais é minimamente equilibrado ou completo para poder representar "o oral". Senão vejamos, através de uma breve panorâmica deste:

As convenções e a informação nos corpos orais transcritos (ou nos corpos que incluem partes de oral transcrito ou teatro) variam entre dois extremos: Por um lado, absolutamente nenhuma informação para além de uma divisão grosseira entre géneros orais, como é o caso dos corpos Borba-Ramsey (ECI-EBR), que marca apenas "oral", do Corpus Brasileiro (CBRAS), cuja divisão está na tabela 1, e da Coleção Dourada do HAREM (CDHAREM), que como oral só tem o género "entrevista". Em todos estes casos, não temos hipótese de saber quem foi o entrevistado ou falante, nem quando a entrevista foi feita.

fb	Política	Debates de TV
fc	Política	Pronunciamentos do presidente
fd	Política	Sessões do congresso
fe	Jornalismo	Entrevistas
fa	Esporte	Narração de jogos de futebol

Tabela 1: distribuição dos géneros orais no Corpus Brasileiro (Berber Sardinha et al., 2008).

Depois, temos dois corpos constituídos por entrevistas, mas muito diferentes: o corpo do Museu da Pessoa (lado português) foi transcrito por alunos sem ter obedecido a quaisquer normas linguísticas, como verificámos ao fazer uma revisão aturada (Taveira & Santos, 2015). A

3 Ver, para mais informações, o sítio do AC/DC, http://www.linguateca.pt/ACDC/. No resto do artigo, os indicadores em maiúsculas referem-se aos nomes dos corpos no AC/DC.

obtenção de dados extralinguísticos foi feita exclusivamente através do conteúdo das entrevistas, de que não se conhece a data exata. Embora o sítio do Museu da Pessoa brasileiro tenha mais informação sobre os entrevistados, também não conseguimos apurar a data das entrevistas. Já o corpo do projeto Diáspora (DIASPORA-TL-PT) foi coligido por linguistas e existem dados sociométricos sobre os entrevistados, mas o seu objetivo não era o estudo da linguagem oral em si (Goglia & Afonso, 2012).

Finalmente, o corpo C-ORAL-BRASIL, de fala espontânea, foi realmente compilado para estudar o oral, o que significa que muitas outras questões foram acauteladas (Raso & Mello, 2012). Paradoxalmente, e devido à transcrição não ter sido normalizada, isto dificulta o processo de comparação com os outros corpos.

No caso do teatro, que também consideramos oral para efeitos de quantificação, o problema é o oposto: como imaginar a forma de falar a partir do texto? As únicas e geralmente poucas indicações cénicas não cerceiam o poder interpretador de cada ator, que, aliás, pode adicionar muito da sua lavra. Mesmo assim, os dramaturgos com certeza imaginam as suas personagens a falar, embora saibamos que a própria linguagem dramática, quando representada, é diferente do português natural.

Apresentamos assim o panorama do oral a que temos acesso, na figura 1 (direita) com os tamanhos respetivos associados na legenda.

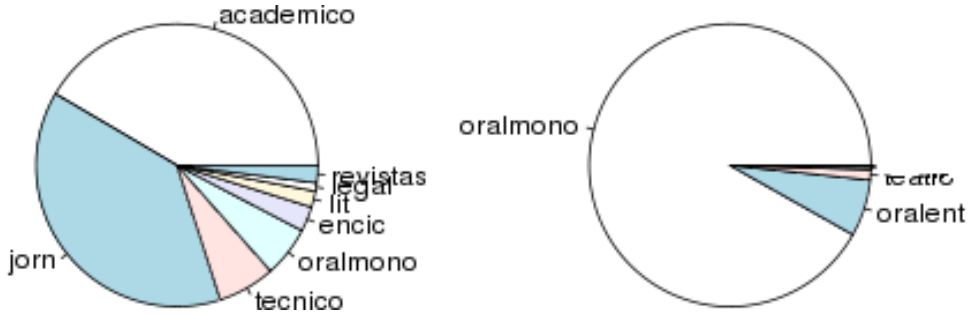

Figura 1: distribuição dos géneros geral, e só orais, na Gramateca: entrevista: 5.559.594, futebol: 86.466, conversa: 292.499, teatro: 1.005.893, debate: 37.317, mono:78.543.738.

Do cotejo das duas figuras, repare-se que a única porção do oral que tem suficiente relevo para figurar no lado esquerdo é a do monólogo... Para ser mais fácil a identificação de todos os géneros presentes, apresenta-se também a lista noutro formato na Figura 2.

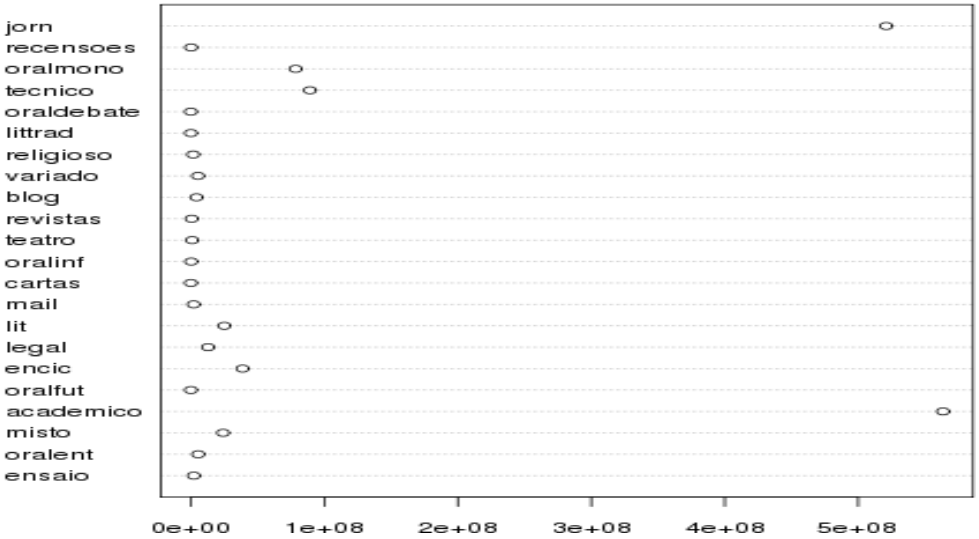

Figura 2: distribuição dos géneros na Gramateca, noutro formato.

2. Alguns estudos preliminares e sua problematização

A primeira pergunta que faz sentido fazer (veja-se Freitas & Santos, 2015) tem a ver com a presença do destinatário do texto, e do autor: Fala-se e escreve-se para alguém, mas quão presente está esse alguém num texto? E, quando se fala muito de si próprio, fala-se menos ao outro, ou textos pessoais também têm muita menção a uma segunda pessoa?

Este tipo de perguntas são mais fáceis de formular do que responder, por várias razões: A primeira é como contar: por número de palavras? Por vez (admitindo que, numa conversa, cada pessoa fala por sua vez)? Ou por mudança de assunto? E interessa simplesmente a frequência num texto, ou a distribuição das menções ao longo do texto?

A segunda questão é que nem sempre os usos de primeira ou segunda pessoa se referem aos participantes no texto: obviamente, basta haver discurso direto para essa hipótese ser falsa. Além disso, em certo tipo de linguagem oral (popular), é frequente usar bordões do tipo *Não acha?*, *Estás a ver?*, *Estás a perceber?* que são tentativas de confirmação de que o ouvinte está a ouvir e não referências ao mesmo.

Seja como for, com estes problemas em mente, vamos começar por tentar – até para mostrar outros problemas que surgem – medir a segunda pessoa em português. “Segunda pessoa” não quer, naturalmente, dizer segunda pessoa gramatical: em português a terceira pessoa gramatical é a mais usada

para a segunda pessoa semântica, desde as formas de respeito *o senhor* e *a senhora* até ao uso generalizado de *você* no Brasil, passando pelo uso raro de *tu* escrito em Portugal, a não ser para pessoas próximas ou entre jovens.

Na tabela 2, apresentamos os números dos pronomes pessoais de segunda pessoa *tu* (*te/ti*) e *você*, *vós* e *vocês* no Museu da Pessoa.

(lema)	total	BR	PT
tu	948	769	179
você	9473	9206	261
vós	50	1	49

Tabela 2: Distribuição dos pronomes de segunda pessoa no corpo Museu da Pessoa.

A primeira observação é a de que não se pode fazer uma comparação direta das entrevistas portuguesas e brasileiras (ou das timorenses), visto que a própria dinâmica das entrevistas é muito diferente: cf. a Figura 3 sobre o tamanho das entrevistas em palavras e em perguntas.

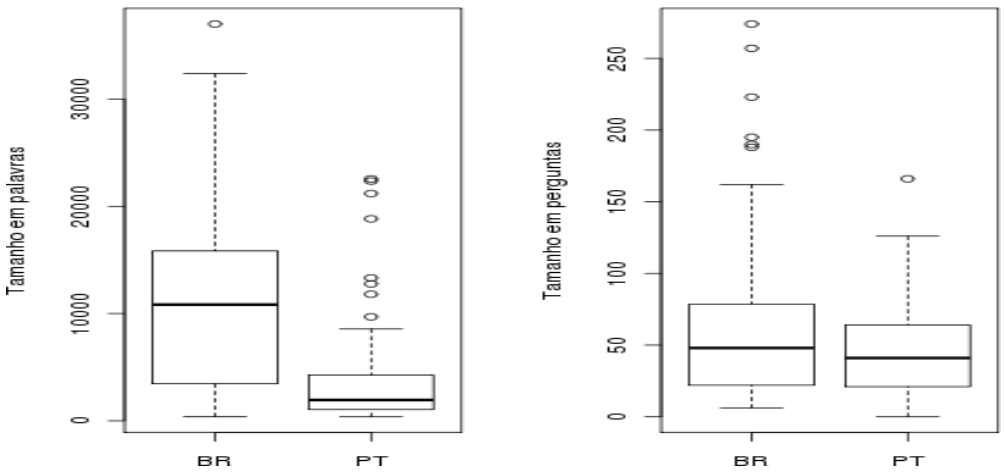

Figura 3: Distribuição do tamanho dos três corpos de entrevistas.

Compreendemos assim que não se podem fazer generalizações sobre a pessoa com base apenas neste tipo de entrevistas: Devido à relação de

respeito (muitas vezes mútuo, por razões diferentes[4]) entre o entrevistado e o entrevistador, o tratamento por *tu* chega a ser mais frequente nas entrevistas brasileiras [5] do que nas portuguesas! Por outro lado, sendo entrevistas sobre a vida de uma pessoa, espera-se que os entrevistados falem sobre si próprios – e ou sobre o meio em que vivem e viveram. Por essa razão, faz sentido comparar a proporção relativa de primeira e segunda pessoa nestas entrevistas, em que, para sermos mais fiéis à língua, incluímos *a gente* como primeira pessoa do plural e *o/a senhor/a* como segunda do singular. A figura 4, dizendo respeito apenas às entrevistas do Museu da Pessoa, compara a frequência da primeira com a segunda pessoa; e a frequência do plural e do singular na primeira pessoa.

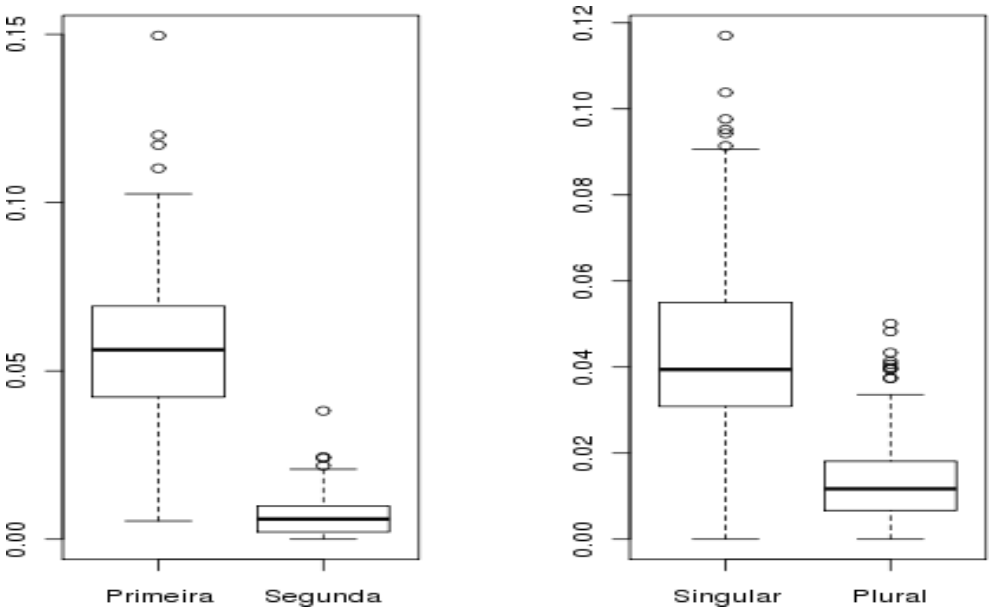

Figura 4: Primeira ou segunda pessoa; na primeira: singular ou plural, no Museu da Pessoa

Comparemos agora as figuras obtidas com as do Corpus Brasileiro, em todos os géneros, na figura 5. Significativo é o facto de os contos (eb) serem um dos géneros com mais incidência de primeira pessoa, ultrapassando o

4 Salvo raras exceções, as entrevistas portuguesas são conduzidas por um universitário jovem a um homem ou mulher idoso, daí o respeito dos entrevistadores perante a idade, e o respeito dos entrevistados perante uma condição socio-cultural geralmente muito superior.

5 Repare-se que geralmente é um tratamento misto, em que coocorrem *você* e *te* numa mesma interpelação/frase.

oral exceto nas entrevistas (fe), e terem uma alta frequência relativa de *você* ou segunda pessoa. Mais uma vez inesperadamente, o género com mais frequência na segunda pessoa são os trechos bíblicos (ev), com *tu*.

Estas pequenas incursões na contagem da pessoa permitem-nos intuir que a comparação entre oral e escrito tem de ser mediada pelo género. Biber (1988) sugeriu sete fatores necessários para distinguir os vários géneros, nenhum dos quais correlacionado com o contraste entre falado e escrito. Contudo, Biber menciona que os géneros escritos aceitam maior variação do que os orais, sendo necessária "considerable research into the range of speech situations and the functions of linguistic features before attempting a macroscopic analysis" (pág. 205).

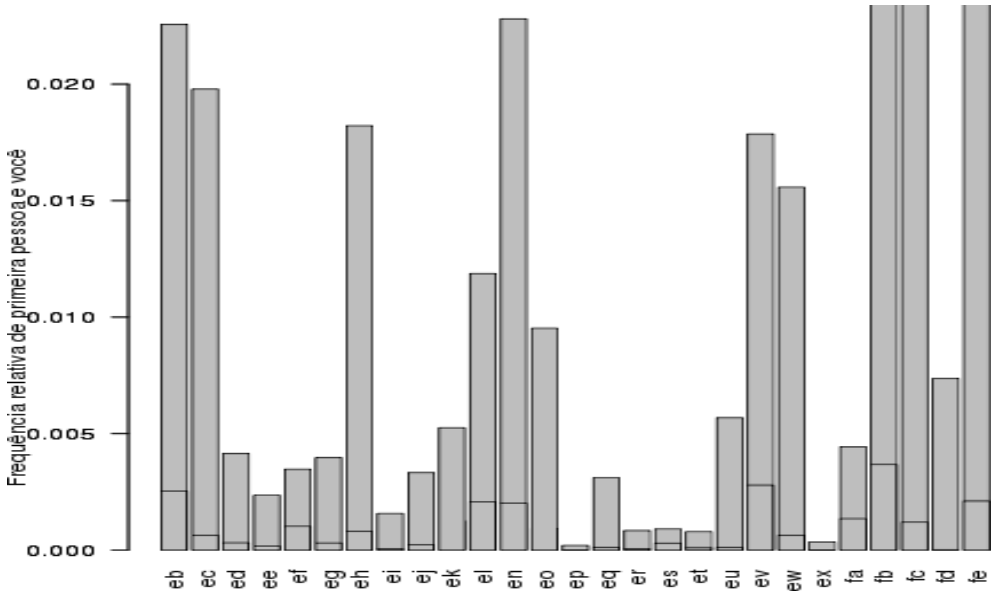

Figura 5: Distribuição de você *e de primeira pessoa no Corpus Brasileiro.*

E como nos dirigimos aos outros? Há tantas maneiras de representar o outro em português (por exemplo, um presidente não trata os seus concidadãos por *vocês*, nem um deputado se dirige ao presidente da assembleia por *tu*) que, dependendo do género (como aliás vimos acima na Bíblia), há formas diferentes esperadas e aceitáveis. Não parece portanto defensável comparar através de uma mesma procura, sem ser mediada pelo género, menção ao outro.

Pode-se, sim, indagar o que se pretende quando se usa a primeira pessoa do plural: contrastar, ou incluir? "Nós inclusivo", ou "nós exclusivo"? (Em

português do Brasil, autores há que consideram *a gente* inclusivo e *nós* exclusivo.[6]) De qualquer forma, concretizando: um presidente a falar na primeira pessoa do plural para o país inteiro, presumindo que *nós* se refere a todo o país, usaria esse método para exortar, ou para descrever? Ao contrário do que o leitor incauto possa pensar, esta pergunta é mais fácil de responder com corpos porque se reflete no modo gramatical: conjuntivo para exortar, mandar, desejar, e indicativo para descrever.

Figura 6: Proporção de conjuntivo e de indicativo por género

Vemos na Figura 6 que o caso de maior frequência de exortação é o do teatro, e da literatura em geral, enquanto que as entrevistas são o género com menor. Aparentemente os discursos monologados têm tanta proporção de conjuntivo na primeira pessoa do plural como os textos jornalísticos. Aqui conviria provalmente distinguir mais finamente que tipos de textos.

Uma carecterística sintática cuja frequência muito se tem discutido é a da passiva, por ser tradicionalmente considerada, por um lado, uma forma mais

---

6 Episódio verídico: Duas portuguesas perguntam a um brasileiro, com quem tinham estado a jantar na véspera*: -- J., a que restaurante nós fomos ontem? -- Não sei. -- Então, você estava conosco! -- Ah, pensei "vocês", não eu também/a gente*. Veja-se Castilho (2010) para referências sobre estas complexidades da pessoa em português.

elaborada de expressão e, por outro, a maneira por excelência de focar em factos (texto "objetivo") despromovendo os agentes da ação. Em Santos (2014), mostrei que a quantificação da passiva traz vários problemas de definição e de operacionalização. No estudo que apresentamos aqui, considerámos como passiva todas as frases com auxiliar *ser*, *estar* e *ficar*, embora indiscutivelmente as diferenças na frequência destes auxiliares contribuam também para a definição de género. A figura 7 compara os diferentes géneros através da proporção de passivas por oração (aproximada por número de verbos), e confirma que as passivas analíticas são, de facto, consideravelmente mais raras no oral.

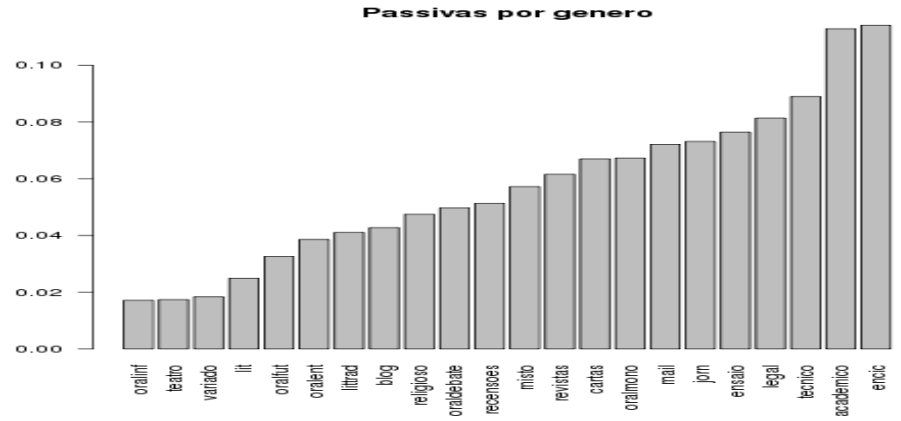

Figura 7: Proporção da passiva por número de orações, nos diferentes géneros

Contudo, algumas surpresas se nos deparam, devidas à pouca representatividade dos textos de alguns géneros: por exemplo, a relativamente alta frequência da passiva nas cartas deve-se ao facto de a correspondência nos nossos corpos ser ou comercial, provinda do corpo NILC, ou de cartas de leitores publicadas em jornais, com um nível de língua elevado. Ainda mais gritante é a constatação de que o correio eletrónico é representado apenas por mensagens não endereçadas, publicitárias, portanto (corpo CoNE), ou constantes duma lista dedicada a conferências sobre bibliotecas (corpo ANCIB). Ambas as fontes são muito diferentes do género informal que associamos ao "email" típico.

Outra pergunta legítima, agora no campo do léxico, é se o oral evidencia mais ou menos palavras de emoção. Existem argumentos óbvios para ambas as posições: -- não, porque a expressão e a entoação chegam; -- sim, porque o oral é mais pessoal. Visto que os corpos foram sujeitos a uma anotação automática de emoção (Mota e Santos, 2014), ver também Maia e Santos (2012), a Figura 8 ilustra uma primeira pesquisa.

Comparando corpos...

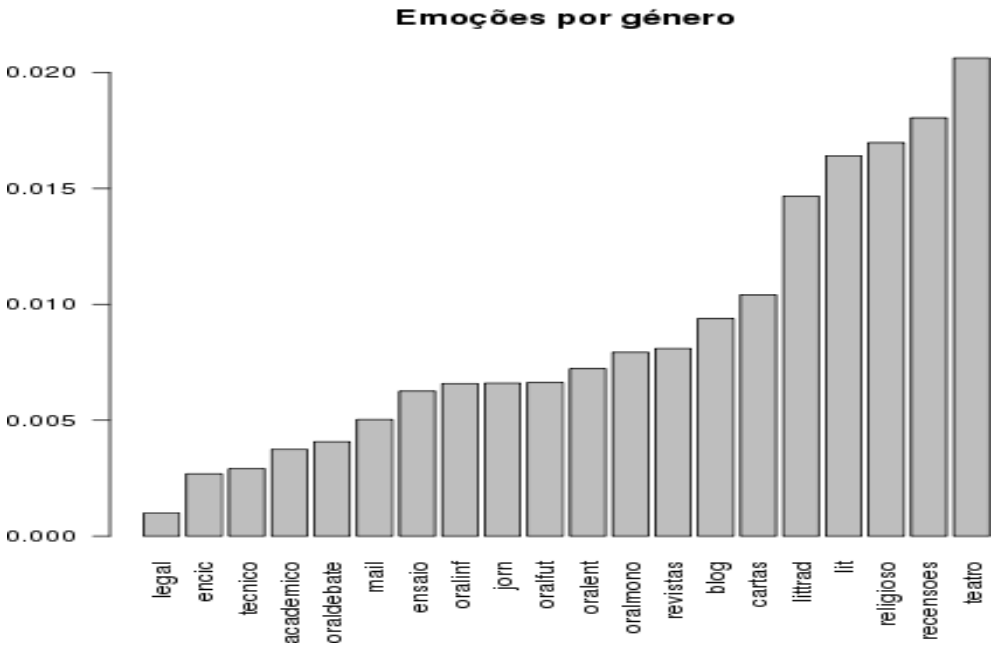

Figura 8: Proporção das emoções por número de palavras, nos diferentes géneros

A resposta é inconclusiva em relação à dicotomia escrito-oral: se nos textos legais e enciclopédicos há um número mínimo de emoções e isso não surpreende, é no teatro e nas recensões (recensões feitas na internete por alunos, e não por críticos literários, Freitas et al. 2012) que as emoções abundam mais, enquanto que a literatura contém muito mais linguagem emocional do que o (resto do) oral. Mais uma vez, o género parece representar um papel importante.

Mas continua a ser necessário sermos críticos: poe exemplo, olhando com mais atenção para as "emoções" nas cartas comerciais, estas são muito menos do que os números automáticos podem fazer crer: de facto, o que acontece é que são palavras de emoção usadas noutro sentido, tão próprio do texto legal ou comercial: *apreciar* não tem a ver com amor mas sim com *preço*, *reconhecer* não se refere a um sentimento de gratidão ou reconhecimento mas a um processo de identificação, *confiar* significa depositar, e aí por adiante... Da mesma forma, olhando para as emoções nas recensões temos de compreender que estas contêm também o enredo.

Tentando aguçar a perspetiva, cruzei então a passiva e as emoções, visto que certo tipo de emoções são geralmente apresentadas na passiva: *Fiquei*

maravilhada (ou *assustada*), *estou espantada* (ou *aborrecida*)... [7] Se concordarmos que a percentagem de casos de passiva nas emoções é indicador de um tipo de linguagem, e que a maior parte da linguagem oral e informal não tem passivas mas contém mais emoções, a figura 9, que ilustra a percentagem de "passivas" relativas a um verbo de emoção, passa a fazer mais sentido. Mas é importante salientar que comparar percentagens, como na figura 9, é enganador quando os números absolutos são muito diferentes.[8]

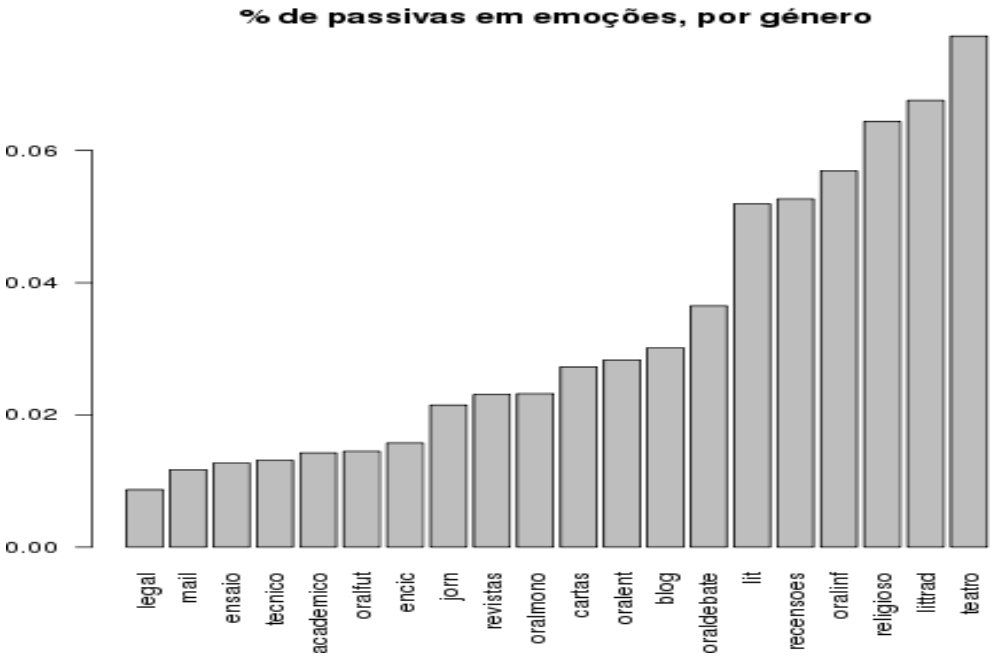

Figura 9: Proporção de emoções em passiva, nos diferentes géneros

Finalmente, debrucemo-nos sobre a palavra *então*, que é usada na oralidade com muitos sentidos discursivos (ver Lopes, 2006), mas que também é usada na escrita como advérbio anafórico. A figura 10 realça a

7 De notar que estes exemplos demonstram claramente que "passiva" não é uma categoria consensual, e muitos certamente consideram que *aborrecida* ou *chateada* são adjetivos plenos tal como *furiosa* ou *triste*. O que me interessa aqui é mostrar que estes "adjetivos" estão marcados como "passiva" nos corpos do AC/DC.

8 Há apenas 672 casos de emoções nas cartas (13 em 477 passivas), comparado com 20.745 casos (160 em 2070 passivas) no teatro.

oralidade desta palavra, que aparece mais frequentemente nos géneros orais, com destaque para a entrevista (brasileira?).[9] Outro estudo que deverá seguir é a diferença do *então* em termos da posição na frase: inicial no oral, seguindo preposição no jornalístico.

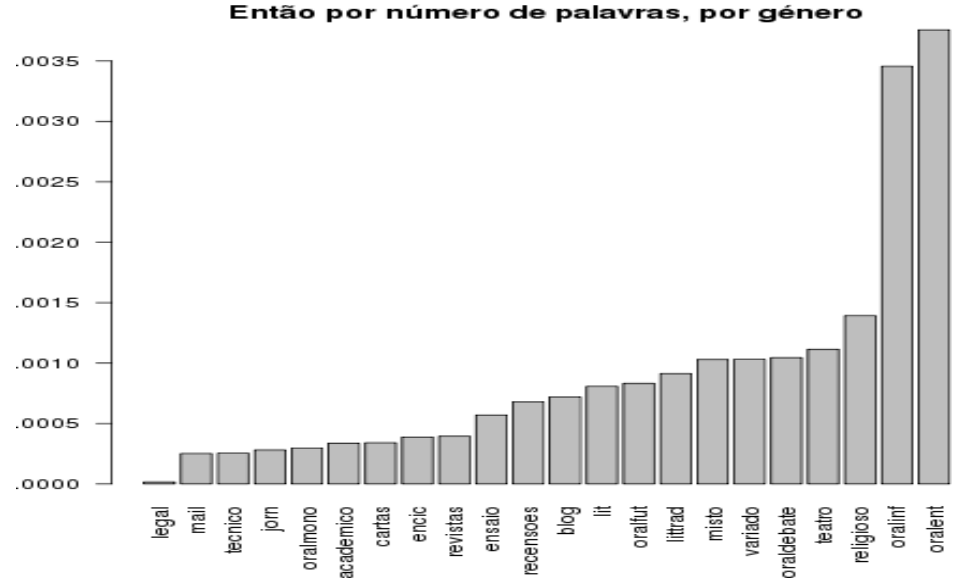

Figura 10: Frequência da palavra então*, nos diferentes géneros*

3. CONCLUSÃO

Embora ainda estejamos nos primórdios de uma caracterização quantitativa adequada dos corpos da Gramateca, pensamos ter ilustrado as suas potencialidades neste assunto tão complexo da comparação entre a língua oral e a escrita em português. A conclusão mais geral é a de que quaisquer tendências preliminares observadas em grandes quantidades de texto têm de ser subsequentemente passadas pela lupa da investigação e interpretação do linguista, chegando por vezes até à interpretação detalhada de cada linha da concordância. Para chegar a conclusões definidas é preciso conhecer bem os materiais sobre os quais se funda a pesquisa, dando lugar a um diálogo de refinamento das hipóteses iniciais, o que é facilitado pela infra-estrutura da Gramateca. Em particular, estabelecemos que:

9 Um dado interessante é que, se nas entrevistas do Museu da Pessoa a proporção de *então* pesa significativamente para o lado brasileiro, com PT: 1,2 por mil, BR: 5,6 por mil; no CONDIV, jornalístico, tal não acontece de todo: PT: 420 por milhão, BR 340 por milhão.

- o modo verbal (o conjuntivo vs. o indicativo) pode ser indicador de interação social (expressão de sentimentos e exortação a ações) em português
- a pessoa gramatical (e as formas de tratamento) são relevantes para caracterizar o tipo de interação
- certo tipo de fenómenos sintáticos (como a passiva), lexicais (como o uso de palavras de emoção) e discursivos (como o uso de *então*) podem ser caracterizadores de géneros escritos ou orais

Pensamos ter aduzido suficiente justificação para a que a separação entre oral e escrito deva ser mediada pelo estudo de géneros orais e escritos.

Agradecimentos

A Gramateca existe no âmbito da Linguateca, co-financiada desde 2000 pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN; e de 2009 a 2011 pela FCT e pela FCCN. Estou grata ao RCS da Universidade de Oslo pelo apoio técnico e ao ILOS pelo semestre sabático em 2014; e a Laura Alvaréz e Camilla Bardel pelo convite para participar no GCSP 2014.

BIBLIOGRAPHY

BENNETT Karen. (2010), Academic Discourse in Portugal: A whole different ballgame?, *Journal of English for Academic Purposes*, 9 (1), 21-32.

BERBER SARDINHA, Tony, MOREIRA FILHO, J. L., & ALAMBERT, E. (2008), O Corpus Brasileiro. Comunicação apresentada em VII Encontro de Lingüística de Corpus, Unesp, São José do Rio Preto, SP, 6 e 7 de novembro de 2008.

BIBER Douglas. (1988), *Variation across speech and writing*, Cambridge University Press.

BIBER Douglas, JOHANSSON Stig, LEECH Geoffrey, CONRAD Susan & FINEGAN E. (1999), *The Longman grammar of spoken and written English*. London: Longman.

BIBER Douglas & GRAY Bethany. (2010), Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *J. of English for Academic Purposes*, 9, 1, 1-82.

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

CASTILHO Ataliba T. de. (2010), *Nova Gramática do Português Brasileiro*, Contexto.

FREITAS, Cláudia, MOTTA Eduardo, MILIDIÚ Ruy Luiz & CÉSAR Juliana. (2012), Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros. In *XI Encontro de Linguística de Corpus - ELC 2012* 13-15 de Setembro

FREITAS Cláudia & SANTOS Diana. (2015), Blogs, Amazônia e a Floresta Sintá(c)tica: um corpus de um novo gênero?, in Ana Maria T. Ibaños, Lívia Pretto Mottin, Simone Sarmento & Tony Berber Sardinha (orgs.), *Pesquisas e Perspectivas em Linguística de Corpus*, Campinas, São Paulo: Mercado de Letras, 123-150.

GALVES Antonio, GALVES Charlotte, GARCÍA Jesús E., GARCIA Nancy L. & LEONARDI Florencia. (2013), Context Tree selection and linguistic rhythm retrieval from written texts, *The Annals of Applied Statistics*.

GOGLIA Francesco & AFONSO Susana. (2012), Multilingualism and Language Maintenance in the East Timorese Diaspora in Portugal, *Ellipsis (Journal of the American Portuguese Studies Association)*, 10, 97-123.

LOPES Ana C. M & AMARAL Patrícia. (2006), From time to discourse monitoring: agora and então in European Portuguese, *Belgian Journal of Linguistics*, 20, 3 - 18.

MAIA Belinda & SANTOS Diana. (2012), "Who's afraid of ... what?" -- in English and Portuguese. *Aspects of corpus linguistics: compilation, annotation, analysis*, ed. by Signe Oksefjell Ebeling, Jarle Ebeling & Hilde Hasselgård. (Studies in variation, contact and change in English 12). Helsinki: Research Unit for Variation, Contacts, and Change in English.

MOTA Cristina & SANTOS Diana. (2014), Emotions in natural language: a broad-coverage perspective, http://www.linguateca.pt/acesso/EmotionsBC.pdf

RASO Tommaso & MELO Heliana (eds.) (2012), *C-oral-Brasil I Corpus de referência do português brasileiro falado informal*, Belo Horizonte: Editora UFMG.

SANTOS Diana. (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, 113-128.

SANTOS Diana. (2014a), Podemos contar com as contas?, in Sandra Aluísio & Stella Tagnin (eds.), *New language technologies and linguistic research: a two-way road*, Cambridge Scholars Publishing, 194-213.

SANTOS Diana. (2014b), Gramateca: corpus-based grammar of Portuguese, in Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A.S. Pardo & Maria das Graças Volpe Nunes (eds.), *PROPOR 2014*, LNAI 8775. Springer, 214-219.

SANTOS Diana & BICK Eckhard. (2000), Providing Internet access to Portuguese corpora: the AC/DC project, in Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhauer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, 205-210.

TAVEIRA Paula & SANTOS Diana. (2015), Ensaio sobre a revisão da oralidade. Em apreciação.