Blogs, Amazônia e a Floresta Sintá(c)tica: um corpus de um novo gênero?

Cláudia Freitas (Linguateca-FCCN / PUC-Rio)
Diana Santos (Universidade de Oslo / Linguateca-FCCN)

Resumo: Simplificadamente, podemos entender gêneros textuais como processos sociais, dinâmicos e interativos que se realizam na ação verbal. O foco da definição está, normalmente, na função. No entanto, a dimensão técnica – o suporte –, embora raramente explicitada, também é tomada como critério na descrição dos gêneros textuais. Com a emergência de novas tecnologias associadas à comunicação, aparecem novos desafios para o estudo dos gêneros, e do espaço ocupado pelo suporte nesses novos espaços. Nesse contexto aparecem os blogs: são um gênero novo, um gênero híbrido, um supergênero, gênero e suporte, simultaneamente, ou apenas um suporte? Embora seja crescente o interesse na compilação de novos corpos para a língua portuguesa, notamos que, em geral, predominam corpos compostos por textos de jornal. Nesse artigo, apresentamos a Amazônia, um corpo de blogs, publicamente disponível, criado pela Linguateca no âmbito do projeto Floresta Sintática. O objetivo do presente artigo, portanto, é duplo: (i) contribuir para uma descrição deste tipo de texto/gênero eletrônico; (ii) apresentar a Amazônia, um corpo composto por textos de um blog coletivo brasileiro, comparando-a com outro material.

Palavras-chave: blogs; gêneros textuais; floresta sintática

Abstract: Briefly, we can view genre as the indication of social, dynamic and interactive processes that take place in verbal communication. Although the definitions usually favor the “functional” aspect, namely what is the communicative function of the text, the technical aspect – the medium – is also taken as a criterion in describing textual genres. In this context, we ask how blogs should be characterized: are they a new genre, a hybrid genre, a supergenre? Or do they simultaneously reflect genre and medium? In this paper, we analyse Amazônia, a publicly available corpus made of blogs, created in the scope of the Floresta Sintá(c)tica project. The goals of the paper are twofold: (i) to contribute to a characterization of this kind of text; and (ii) to provide a presentation of Amazônia featuring some studies on its style and content.

Keywords: blogs, textual genres, treebanks

1. Introdução

O estudo dos gêneros textuais tem recebido atenção crescente por parte da Linguística, tendo-se cunhado, inclusive, o termo “Linguística dos Gêneros” (Ciapuscio, 2009; Adamzik, 2000 apud Ciapuscio, 2009). Usa-se o termo gênero não apenas no âmbito dos estudos de gênero, mas também, e cada vez mais, no âmbito

dos estudos de corpus (ou “corpo”, plural “corpos”, como defendemos em Santos, 2008). E, como frequentemente acontece com termos de ampla utilização, é natural que o seu significado tenda a variar.

Do lado dos estudos de gênero podemos, simplificadamente, entender gêneros textuais como processos sociais, dinâmicos e interativos que se realizam na ação verbal. De modo complementar, um gênero textual também pode ser visto como uma maneira de realizar objetivos por meio da linguagem (Marcuschi, 2004, 2008; Miller, 1984, Eggins & Martin, 1997). Neste caso, o foco da definição está, principalmente, na sua função, nos objetivos pretendidos com uma determinada ação verbal. Assim, a distinção entre gêneros considera, principalmente, a dimensão de funcionalidade, e a tipicidade de um gênero vem com suas características funcionais e organização retórica. Estão em jogo, principalmente, critérios como objetivo, função e comunidade discursiva.

No entanto, a dimensão técnica – o suporte –, embora raramente explicitada, também é tomada como critério na descrição dos gêneros textuais: “Os gêneros textuais (...) definidos por composições funcionais, objetivos enunciativos e estilos concretamente realizados na integração de forças históricas, sociais, institucionais e *técnicas*” (Marcuschi 2008:155, grifo nosso).

Quando se assume que um bilhete pode ser “transformado” em recado, caso esteja em papel ou secretária eletrônica, quando se postula que o meio é capaz de interferir de maneira tão direta na classificação de gêneros, e quando se tem a emergência de novas tecnologias associadas à comunicação e à escrita, aparecem novos desafios para o estudo dos gêneros, e do espaço ocupado pelo suporte nesses novos contextos. A definição do status dos blogs é um desses desafios: são um gênero novo, como sustentam Marcuschi (2004, 2008), Miller & Shepherd (2004), um gênero híbrido (Herring et al, 2005), um gênero da comunicação mediada por computador, que por sua vez contém subgêneros como os diários pessoais ou blogs corporativos, apenas um suporte ou suporte e gênero, simultaneamente?

Se, inicialmente, blogs surgiram com o propósito de relatar experiências pessoais, atualizando os diários pessoais, cada vez mais eles exercem diferentes funções: para jornalistas, podem servir como espaços alternativos de publicação; para empresas (mas não somente), mais um espaço de divulgação; para boa parte dos blogueiros, um espaço de interação, expressão pessoal e compartilhamento de ideias. Tratam dos conteúdos mais diversos, podem ser escritos em diferentes registros, e compreender textos em gêneros variados – de ensaios a receitas culinárias, de resenhas e entrevistas a relatos de viagens e de eventos, por exemplo. Podem ser de um único autor ou escritos coletivamente, e seus escritores têm uma formação diversificada, o que também contribui para a diversidade dos estilos de escrita. Há, no entanto, pouca investigação até o momento sobre como de fato são ou quais as funções mais comuns dos blogs (Herring et al., 2005).

Blogs caracterizam-se principalmente pela subjetividade. São, simultaneamente, públicos e pessoais. Contêm relatos pessoais que expressam a opinião com relação a "objetos" como empresas, eventos, pessoas e coisas. Por isso, textos veiculados em blogs interessam tanto à descrição de práticas linguísticas, sociais e culturais (Herring et al. 2005; Miller & Shepherd 2004; Blood, 2000) como `a área de detecção de opinião em textos (Andreevskaia et al, 2007, Leshed &Kaye, 2006) e extração da informação (Barbosa et al, 2009), por exemplo.

Ao lado do email, chat e homepage, blogs e microblogs integram o grupo dos chamados gêneros digitais, ou gêneros eletrônicos, ou ainda gêneros que integram a "comunicação mediada por computador" (CMC), aqueles que surgiram em decorrência da tecnologia digital em ambientes virtuais (e, de forma subjacente, já aparece a ideia de que todos esses "textos" são, de fato, "gêneros", o que será discutido na seção 2), cuja análise se justificaria principalmente por (i) o amplo desenvolvimento e uso cada vez mais generalizado; (ii) as características formais e funcionais; (iii) a possibilidade de revisão de nossa relação com a oralidade e com a escrita (Marcuschi, 2005).

A CMC caracteriza-se pela falta de retorno simultâneo enquanto escrevemos – por mais que a interação se dê por meio de mensagens "instantâneas", a interação é sequencial. Não temos retorno em termos de movimentos de cabeça, olhares, assim como nosso interlocutor não pode se apoiar em pistas como tom de voz ou expressão facial. Como bem nota Crystal (2004), nós escrevemos digitando uma tecla por vez, mas nosso interlocutor não recebe um caractere por vez. Ele só recebe nossa mensagem quando damos o comando "enviar", o que significa que a mensagem é transmitida e recebida como uma unidade. Não há, até o momento, como reagir à mensagem enquanto é digitada – não é possível um *ahã*, ou qualquer outro tipo de reação, simultâneo. Com isso, os participantes desse tipo de interação verbal passaram a introduzir "sinais" que têm por objetivo minimizar a ausência de algumas dessas pistas durante a interação. Consequentemente, temos uma escrita que busca se aproximar cada vez mais da fala, incluindo-se no "texto" marcadores verbais como o próprio *ahã*, além de outros como *hmmm*, *hã?*, mas também as letras repetidas (*claaaro*; *oooooi*); marcadores para expressar ênfase (como uso de maiúsculas ou de símbolos como " *o *verdadeiro* problema*") além dos emoticons (Crystal, 2004). Uma caracterização da linguagem dos blogs forneceria mais uma pista sobre esse novo modo de agir.

Acreditamos que essa breve discussão seja suficiente para ilustrar a relevância de um corpo de blogs. Embora seja crescente o interesse na compilação de novos corpos para a língua portuguesa, notamos que, em geral, predominam corpos compostos por textos de jornal. Com relação a textos criados especificamente para ambientes digitais – como os blogs –, e no que pese a relativa facilidade na obtenção de dados de blogs – ou de outros textos veiculados na internet – e o interesse pela linguagem utilizada nos textos da CMC, temos conhecimento apenas

do Corpo ANCIB, com cerca de 81.000 frases e que corresponde a mensagens de correio eletrônico da lista ANCIB[1], cujo conteúdo é bastante previsível e institucional (anúncios de conferências, majoritariamente), e da Coleção Dourada do Segundo HAREM (Carvalho et al, 2008) que inclui, dentre outros tipos de textos, blogs e textos da Wikipédia.

Nesse contexto, apresentamos a Amazônia, um corpo de blogs, publicamente disponível, criado pela Linguateca no âmbito do projeto Floresta Sintática (Afonso et al., 2002). A Amazônia contém 4.6 milhões de palavras da variante brasileira do Português, analisadas morfossintaticamente pelo analisador PALAVRAS (Bick, 2000).

O objetivo do presente artigo, portanto, é duplo: (i) contribuir para uma descrição deste tipo de texto/gênero eletrônico, que, devido à crescente produção, facilidade de obtenção e ausência de manuais/liberdade de escrita, é uma amostra rica da língua em funcionamento; (ii) apresentar a Amazônia, um corpo composto por textos de um blog coletivo.

2. Blogs, gênero e suporte

Blogs devem sua existência à internet, que tem como características principais a rapidez, interação e fluidez: usuários exploram as possibilidades de expressão, introduzem combinações novas de elementos e reagem, à sua maneira, aos desenvolvimentos tecnológicos (Crystal, 2004). Em Harring et al. (2005), por exemplo, uma das características atribuída aos blogs é serem mantidos por uma única pessoa, assim como as páginas pessoais. Hoje em dia, no entanto, é comum blogs serem escritos e mantidos coletivamente. A fluidez da Internet transmuta gêneros existentes, criando alguns novos e mesclando outros. Qual o lugar dos blogs nessa nova "ecologia" dos gêneros? Seriam eles realmente um gênero novo, ou uma mescla?

Antes, porém, de problematizarmos o blog enquanto gênero, é relevante uma tentativa de entendimento do que se entende por gênero, tarefa nada fácil. De fato, a noção de gêneros textuais parece mais um dos terrenos "movediços" da linguística. "Dar nome aos gêneros é algo de enorme complexidade", como afirma Marcuschi (2008: 161), e podem ser vários os critérios utilizados na sua identificação.

Gêneros textuais são abstrações. Enquanto construções sociais e históricas, são objetos linguísticos cuja definição está fora do escopo da linguística, na prática ou ação social. Assim, para pensarmos se blog é ou não gênero, é preciso olhar para fora da língua.

Segundo Bakhtin (1986, p. 60), gêneros são "tipos relativamente estáveis de enunciados". No entanto, a fluidez apontada por Crystal, característica da internet,

põe em xeque a estabilidade mencionada por Bakhtin, ainda que tal estabilidade seja relativa.

Para Miller (1984, seguido de Swalles, 1990), um indicador da existência de um gênero novo seria a atribuição de um nome, reconhecido pela comunidade, para uma dada ação discursiva ou comunicativa. Esse critério, no entanto, não nos ajuda muito, pois temos uma vagueza entre o blog "suporte" (entendido como uma ferramenta de publicação) e o blog tipo de texto – o mesmo acontecendo, aliás, com o jornal.

Marcuschi refere que "(...) para a noção de gênero textual, *predominam* os critérios de padrões comunicativos, ações, propósitos e inserção sociohistórica" (2008:158, grifo meu), ainda que, poucas páginas atrás, admita a importância da dimensão técnica na caracterização dos gêneros. Araújo (2006, apud Travaglia, 2007), aliás, exemplifica de maneira muito clara a interferência do suporte na caracterização do gênero: uma mesma "mensagem" como "Parabéns! Toda a felicidade do mundo, hoje e sempre. Beijos, Cláudia", pode ser um bilhete se um pedaço de papel deixado em cima da mesa, um recado se deixado na secretária eletrônica, um email se na caixa do correio eletrônico, ou mesmo um telegrama. Ou seja, embora o conteúdo seja exatamente o mesmo, que determina a mudança em termos de gênero é o suporte, é o meio em que o conteúdo foi veiculado – o que transforma automaticamente o blog em um novo gênero.

Se o foco da definição (e distinção) entre gêneros está na dimensão da *ação* social, o blog também pode ser entendido como uma *nova forma de ação*, que traz pessoalidade e aproxima interlocutores.

Se, por outro lado, o foco está na dimensão "função", não necessariamente o blog se caracteriza como um gênero novo, pois não há, ali, uma função nova intrínseca (diferentemente da página pessoal (homepage), por exemplo, que seria um gênero eletrônico realmente novo, cuja existência é decorrente da internet). Ou, ainda, pode-se pensar que a função do blog é nova, mas não necessariamente é nova a função dos textos veiculados pelos blogs (artigos, relatos, resenhas, convites etc). E, com isso, já pressupomos também a distinção entre o gênero blog e o suporte blog.

O trabalho de Herring e colegas (2005) busca caracterizar o *gênero* emergente blog, e situá-lo com relação aos demais gêneros convencionais, chamados, pelos autores, de gêneros "off-line", e a conclusão a que chegam é de que blog não é algo fundamentalmente novo nem único, mas um gênero híbrido. Para Blood (2000), Miller & Shepherd (2004) e Marcuschi (2005), o blog é gênero diferente, novo. "*Os blogs têm uma história própria, uma função específica e uma estrutura que os caracteriza como um gênero, embora extremamente variados nas peças textuais que albergam. Hoje são praticados em grande escala e estão fadados a se tornarem cada vez mais populares pelo enorme apelo pessoal."* (Marcuschi, 2005, p.61).

No entanto, ao assumir gênero como uma classificação que privilegia a função, parece difícil sustentar o hibridismo ou inedistismo dos blogs, pois não há nada de novo *nesses* termos; não há novidade no blog enquanto função, mas na facilidade de distribuição e divulgação do seu conteúdo.

Ao tentar definir blog como gênero (ou não) para a caracterização da Amazônia, acabamos por aproximar o trabalho com base em corpos com os estudos do gênero. No entanto, como notam Mauranen (1998) e Berber-Sardinha (2003), a aproximação entre as duas áreas ainda é tímida. "Seria muito benéfico para ambas as áreas se houvesse uma maior aproximação entre elas" (Berber-Sardinha, 2003:2). Ainda no mesmo trabalho, o autor aponta a falta de cuidado na descrição de corpos quanto ao gênero: o BNC, por exemplo, ofereceria um tratamento equivocado de gênero na sua composição (Lee, 1999, apud Berber-Sardinha 2003).

2.1 Corpos e gênero

Assumimos, neste trabalho, que corpo é uma coleção classificada de objetos lingüísticos para uso em Processamento de Linguagem Natural/Linguística Computacional/Linguística; e que a compilação dos textos de um corpo deve ser feita associada a algum objetivo (Santos, 2008).

Com relação à Amazônia, o interesse principal está na compilação de textos veiculados em blogs, escritos por diversos autores e que fugissem ao discurso jornalístico. Como já tínhamos conhecimento do Overblog, optamos por compilar os textos disponíveis na época. Um ponto a lamentar é a inexistência de material com um perfil semelhante na variante portuguesa, mas nada impede que, no futuro, textos de blogs portugueses sejam incorporados.

Com relação aos usos que se dá aos corpos, é possível distingui-los em "estritamente linguísticos", como pesquisas exploratórias e experimentais; e usos aplicados, como a criação de dicionários, ontologias e tesauros; o treino e avaliação de sistemas que lidam com a língua, bem como dos métodos usados por esses sistemas; e o auxílio na a elaboração de material didático, como jogos (Santos, 2008). Amazônia pode, inicialmente, ser útil ao primeiro dos usos descritos acima – a exploração de parte da língua, em especial da linguagem dos blogs. Um estudo exploratório, como o nome indica, "procura coisas interessantes para mais tarde estudar. Colige amostras, surpreende-se (...). Por outras palavras, abre sendas, identifica lugares de interesse (para lá voltar ou para outros lá irem)" (Santos, 2008:49)

No entanto, dados os variados usos que se pode dar a um corpo, e considerando ainda aqueles não previstos, é de extrema importância a documentação cuidadosa deste tipo de recurso, para que seja realmente usável e, os resultados obtidos a partir da sua exploração, comparáveis. (Sobre a relevância da documentação dos corpos, ver Santos, 2008). Uma das dimensões da documentação é o gênero.

Um corpo cuidadosamente descrito quanto ao gênero, além de uma melhor compreensão dos resultados das pesquisas efetuadas no próprio corpo, permite a investigação e descrição detalhada dos padrões que constituem o próprio gênero, tanto em termos de marcas lexicais /gramaticais, quanto em termos de movimentos discursivos – ainda que, normalmente, este último aspecto seja pouco considerado, e a (pertinente) crítica de Mauranen (1998) é de que os atuais corpos eletrônicos dificilmente são apropriados aos estudos dos gêneros, dados os diferentes rumos seguidos pelas duas áreas.

A linguística com base em corpos tem privilegiado a quantidade, buscando cobrir vastas porções da língua. Do lado dos gêneros, a crítica que se faz é que, nesta tentativa, os corpos compilados acabam por ser generalistas, não sendo suficientemente representativos em termos de gênero. Já a teoria dos gêneros vem se desenvolvendo em um viés de pesquisa qualitativo, e a crítica aqui está na ausência de uma tentativa de caracterização abrangente dos gêneros, ou mesmo de uma validação empírica dos gêneros postulados. Claro está, portanto, que ambas as perspectivas só têm a ganhar com o estabelecimento de bases “compatíveis”. E, neste ponto, voltamos ao impasse da seção anterior.

Gêneros são objetos multidimensionais, e podem ser definidos tomando por base sua função, objetivo, mas, também, o suporte – o que interessa aqui.

Como nota Mauranen (1998), a caracterização de gêneros a partir de traços externos como dinamicidade, contextualização, comunidade discursiva, forma e conteúdo (Barkenkotter e Huckin, 1995, apud Mauranen, 1998) é eficaz apenas se estamos diante de gêneros já estabelecidos, em uma atitude de reconhecimento. Quando se está diante de um gênero novo, ou justamente tentando verificar a existência de um gênero, é de pouca utilidade. Do mesmo modo, trabalhos como o de Berber-Sardinha (2003), que buscam a identificação das unidades internas/elementos estruturais em diferentes gêneros, partem de uma definição já consensual de um dado gênero – situação que, tendo em vista a emergência dos “textos em suporte eletrônico”, não é tão estável.

Apresentamos a seguir a Amazônia, bem como uma breve exploração do seu conteúdo, tendo como objetivo auxiliar a elucidação de algumas características textuais capazes de ajudar a pensar o lugar dos blogs quanto ao gênero.

3. A Amazônia e o projeto Floresta Sintá(c)tica

A Amazônia é uma das partes da Floresta Sintá(c)tica (www.linguateca.pt/Floresta), uma colaboração entre a Linguateca e o projeto VISL (Visual Interactive Syntax Learning). "Floresta Sintática" é um conjunto de frases analisadas (morfo)sintaticamente. Como, além da indicação das funções sintáticas, a análise também explicita hierarquicamente informação relativa à estrutura de constituintes,

dizemos que uma frase sintaticamente analisada se parece com uma árvore, donde um conjunto de árvores constitui uma floresta (em inglês, *treebank*).

Uma floresta sintática também pode ser caracterizada pela revisão linguística (humana) das árvores automaticamente analisadas (Afonso et al., 2002). Ainda que sejam óbvios os ganhos com a revisão linguística, a necessidade de revisão humana não é, contudo, entendida como imprescindível para a existência de uma floresta, embora seja altamente desejável.

A riqueza da informação codificada nos corpos da Floresta faz com que esta seja uma valiosa fonte de pesquisa. Graças à rica anotação linguística e à existência do formato de árvore, é possível investigar questões relacionadas a (ou, realizar buscas que envolvam) hierarquia entre constituintes da oração, tipos de enunciado (perguntas, ordens, declarações, exclamações), tipos de oração (finitas, não finitas), funções sintáticas (argumentos e modificadores verbais, nominais, adverbiais; aposto etc) e uma ampla variedade de formas (sintagmas verbais, nominais, preposicionais, adverbiais), além da informação disponível nas chamadas etiquetas secundárias, como lemas e informação morfossintática. Para facilitar o acesso à boa parte desses dados, a Floresta conta com o que chamamos "procuráveis": etiquetas que visam simplificar a busca por estruturas como orações na voz passiva; orações passivas com o –se; orações sem sujeito explícito e sem sujeito formal; orações substantivas, relativas e diversos tipos de orações adverbiais (concessivas, causais, conformativas, temporais etc) e estruturas com quantificadores. O Bosque, por ter sido integralmente revisto, e a Selva, parcialmente revista, possuem informação ainda mais refinada: e etiqueta n-adj para os casos em que achamos difícil decidir entre um substantivo e um nome, e preferimos não tomar posição, deixando o campo aberto à exploração (frases a-c abaixo)

a) Embora seja concorrente respeitado, a Nielsen não representa uma ameaça real », diz Flávio Ferrari, diretor da Ibope Mídia.
b) Os jovens têm se encontrado em Atalaia, principalmente entre quarta-feira e domingo, quando são realizados jogos de voleibol.
c) A pizzaria foi inaugurada pelo Grupo Viena há dois meses e oferece 21 tipos de pizzas no forno à lenha, além de massas especiais, grelhados e saladas.

e as etiquetas N<ARGO; N<ARGS, indicadoras de argumentos de nomes e que correspondem aos casos em que tais argumentos teriam uma função de objeto ou sujeito, respectivamente – e que se opõem aos modificadores do nome (N<) –, e que são uma especificação da etiqueta mais abrangente N<ARG (frases d-f, abaixo).

d) Governador do Estado acusa governo federal dos EUA de não impedir entrada de estrangeiros. (N<ARGS)

e) O economista diz que, para geração de empregos, várias propostas terão que ser analisadas. (N<ARGO)

f) As filas gigantescas à entrada do imponente mausoléu na Praça Vermelha desapareceram. (N<)

Iniciado em 2000, o projeto Floresta Sintá(c)tica tem, atualmente, quatro partes, que diferem quanto ao tipo de texto, quanto ao assunto, quanto ao modo (escrito vs transcrição de fala) e quanto ao grau de revisão linguística, descritas a seguir.

Bosque: totalmente revisto por linguistas, é composto por quase 10.000 frases retiradas dos corpos jornalísticos CETENFolha (parte do corpo NILC/São Carlos, retirado de textos do jornal brasileiro Folha de São Paulo, textos de 1994) e CETEMPúblico (retirados do diário português PÚBLICO, textos de 1991 a 1998). Por ter sido integralmente revisto, é o corpo é mais correto da Floresta, e por isso o mais aconselhado para pesquisas em que não se prioriza tanto a quantidade, mas sim a qualidade dos resultados (decorrentes da análise linguística). Toda a documentação do Bosque está disponível na página da do projeto Floresta Sintá(c)tica, e uma extensa documentação das opções linguísticas subjacentes à anotação está em Freitas & Afonso (2008).

Selva: parcialmente revista, a Selva contém cerca de 300 mil palavras nas variantes portuguesa e brasileira do português. A Selva foi criada para ser um corpo parcialmente revisto, e a parcialidade refere-se não à quantidade de revisão feita, mas sim à qualidade. Algumas características foram linguisticamente revistas, e, portanto, a revisão não foi feita árvore a árvore, mas caso a caso. A Selva se subdivide em 3 partes: Selva Falada, Selva Científica e Selva Literária. A Selva Falada é composta pela transcrição de dois tipos de fala: entrevistas e debates parlamentares. A Selva Literária contém textos literários do final do século XIX e do início do século XX, recolhidos na Wikisource e também textos contemporâneos. A Selva Científica contém uma minuta do Banco Central do Brasil, quatro relatórios do Banco Central Europeu, capítulos de teses (brasileiras e portuguesas) e artigos da Wikipédia sobre assuntos relacionados a ciências como astronomia, biologia, física, química, geografia, geologia, história, linguística, computação e zoologia.

Floresta Virgem: como o nome indica, trata-se de uma floresta virgem, isto é, não revista. A Floresta Virgem é composta de cerca de 95.000 frases (cerca de 1.600.000 palavras) retiradas do início dos corpos CETENFolha (parte do corpo NILC/São Carlos) e CETEMPúblico (retirados do jornal português PÚBLICO).

Amazônia: também virgem, isto é, não revista, a Amazônia contém 4.6 milhões de palavras (cerca de 275 mil frases) retiradas do sítio colaborativo Overmundo, um coletivo virtual que tem como objetivo expressar a produção cultural brasileira, e será detalhada a seguir.

Junto, todo esse material soma cerca de 261 mil frases (6,7 milhões de palavras) sintaticamente analisadas pelo analisador sintático PALAVRAS (Bick, 2000). Todo o material da Floresta Sintá(c)tica, está publicamente disponível, para consulta on line

por meio do projeto AC/DC (Acesso a Corpos/ Disponibilização de Corpos – www.linguateca.pt/ACDC) ou das ferramentas de busca criadas especialmente para ela (Milhafre; Águia e CorpusEye); e também para download a partir da página da Linguateca. Excetuando-se a Amazônia, uma descrição mais detalhada do projeto e dos corpos referidos estão em Afonso et al.(2002), Bick et al. (2007), Freitas et al.(2008), além do material disponível na página de documentação do projeto.

3.1 Descrição da Amazônia

A Amazônia contém os textos da seção OverBlog do sítio colaborativo Overmundo, disponíveis em 30 de Setembro de 2008, num um total de 3889 textos (e cerca de 4.6 milhões de palavras / 275 mil frases) de cerca de 1500 autores diferentes. O Overmundo é voltado para a cultura brasileira e o OverBlog, especificamente, contém "reportagens, entrevistas e críticas sobre cultura do Brasil".

Ser "colaborativo" significa que qualquer um pode se registrar e enviar conteúdos, já que "*Nenhuma equipe de jornalistas, não importa seu tamanho ou competência, consegue cobrir ou filtrar a quantidade cada vez maior de coisas importantes que acontecem pelo país*." O objetivo do Overmundo, segundo os próprios organizadores, é "*servir de canal de expressão, debate e distribuição para a produção cultural do Brasil e de comunidades de brasileiros espalhadas pelo mundo afora tornar-se visível em toda sua diversidade."* Em termos de interesse linguístico, como qualquer um pode colaborar - e a ideia é que participem textos de todo o Brasil - não há (tanta) presença do jargão jornalístico, e tenta-se fugir da "hegemonia linguística" do eixo Rio-São Paulo.

Observando o quadro 1, que apresenta alguns trechos de textos da Amazônia, vemos o quão diversificados estes são em termos de escrita. Nos trechos 1 e 5 temos um registro formal, e uma escrita que pode ser caracterizada como acadêmica. Nos outros trechos, observamos o tipo de escrita mais frequente na internet: informal e aproximada da fala: "mó galera" (maior galera); uso de emoticons e de marcadores típicos da oralidade, como "hmmm", "hehe". Pelos exemplos, já é possível ter uma ideia da diversidade dos textos que compõem a Amazônia. Até que ponto considerá-los todos igualmente "blogs" é impor uma uniformidade talvez inexistente?

(1) "Lacuna esta que, ao longo dos últimos anos, buscou-se suprimir com um conjunto de pesquisas acadêmicas, nos cursos de História, Letras e Comunicação, abordando desde a história dos festivais de música realizados na cidade (como o

(5)"Recentemente, o IBAMA, através do CGPEG (Coordenação Geral de Petróleo e Gás), tem intensificado a pressão para que as empresas implementem projetos ambientais consistentes na região, incluindo os de Educação Ambiental,

histórico Massafeira, realizado em 1979 com a participação de vários artistas locais, materializado em disco duplo) à riqueza lingüística das composições.”	historicamente relegados à última prioridade."
(2) “Podendo ajudar mó galera a pertubar o status quo, vou ficar feliz pra caralho! ! !”	(6)“E a gente não tem hit nenhum em rádio, haha, neguinho não vai nem conhecer”
(3)"acho um grande passo pelo menos trocarmos pra ‘detesto’, que é menos pesado que ‘odeio’, será que conseguimos? ;)"	(7)"Coloque sua cidade pra aparecer!:D Seja a mídia que você quer ver no mundo! Hehehe”
(4) "E, acreditem, essa galera ainda vai dar o que falar .. : D Vê o blog lá: "	(8)“Hummm, não sei... ;)”

Quadro 1: Trechos do material que compõe a Amazônia

Na Amazônia, para cada texto, além do número identificador do extrato, estão disponíveis informações referentes ao título, autor, região do Brasil e a localização (URL) do texto . A figura 1 apresenta a distribuição dos textos por estado do Brasil.

Como mencionado, uma das motivações para a criação da Amazônia foi a possibilidade de criar um corpo com textos que fossem representativos dos diversos estados do Brasil, e não apenas Rio de Janeiro e São Paulo (mais frequentes na maioria dos corpos). Como mostra a figura 1, ainda assim São Paulo e Rio de Janeiro continuam a ser os estados com mais contribuições (716 e 590 textos, respectivamente) seguidos por Minas Gerais (245), Rio Grande do Sul (195), Bahia (149) e Pernambuco (143). Os estados com menos contribuições são Maranhão (15), Amapá (18), Roraima (19) e Rondônia (20). A indicação www refere-se a contribuições cujos autores estavam fora do Brasil e correspondem a 87 contribuição de 13 países (Japão (25), África do Sul (27), Afeganistão (8), Austrália (8), EUA (4), França (3), Portugal (3), Canadá (2), Alemanha (2), Emirados Árabes (2), México (1) , Peru (1) e Argentina (1)). Ou seja, ainda que nossa intenção fosse ter uma amostra maior de textos produzidos por todo o Brasil, e embora tenhamos contribuições de todos os estados, os grandes centros continuam a ser os mais representados, o que também não chega a surpreender.

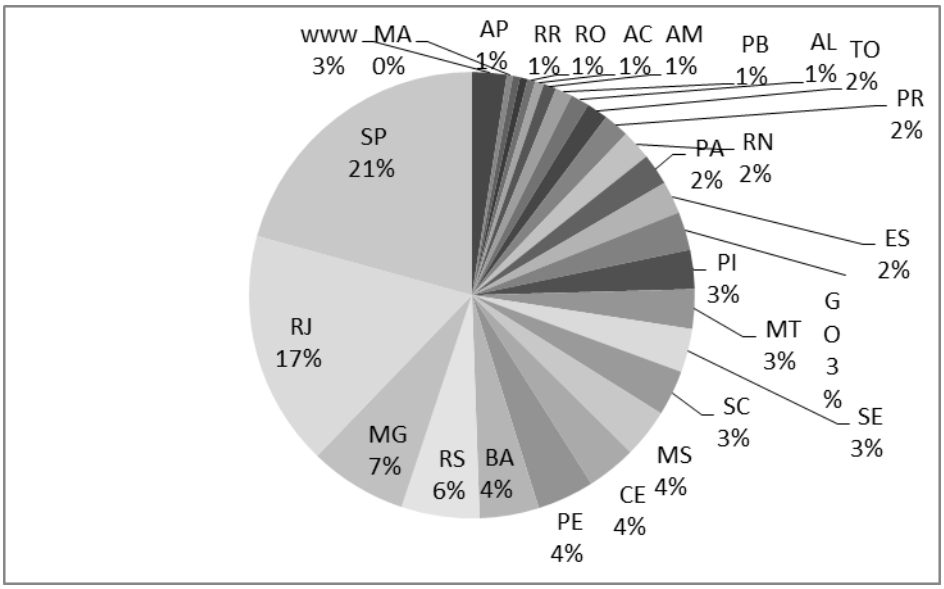

Figura 1: Distribuição dos textos da Amazônia por estado

3.2 Breve exploração da Amazônia

É consenso que os textos da CMC contêm, em maior ou menor grau, indicadores cujo objetivo é evidenciar o caráter interativo e dialógico do texto, minimizando a ausência do interlocutor. É igualmente consensual que os blogs, por mais variados que sejam, têm um forte viés subjetivo. No entanto, os textos do Overblog – "*reportagens*, *entrevistas* e *críticas* sobre cultura do Brasil" – compilados na Amazônia, ainda que integrem um blog, também se aproximam do tipo de texto de um jornal convencional. Até que ponto tais textos apresentam, de fato, diferenças com relação às "*reportagens*, *entrevistas* e *críticas"* (ou *reportagens, entrevistas e resenhas)* convencionais, a ponto de perderem esse traço primeiro de identidade, e devem ser caracterizados como textos de blogs, ou, por outro lado, as diferenças são mínimas, e não justificariam sua caracterização como a de textos de um novo gênero?

Aproveitando, por um lado, toda a informação linguística codificada na Amazônia e, por outro, as possibilidades de pesquisa em corpos oferecidas pela ferramenta de procura em corpos AC/DC (Acesso a corpos/Disponibilização de corpos[2], Costa et al., (2009)), e tendo em vista a influência que a internet exerceria nos textos produzidos, a seguir exploramos brevemente, na Amazônia, marcas linguísticas de interação e de subjetividade, comparando-as com outros corpos disponibilizados pela Linguateca. Dentre os 20 corpos passíveis de busca no AC/DC, foram escolhidos para a comparação com a Amazônia os corpos CHAVE (jornal impresso) e Museu da Pessoa (entrevistas transcritas).

O CHAVE contém os textos completos do jornal português PÚBLICO e do jornal brasileiro Folha de São Paulo, de 1994 e 1995. O interesse na comparação com o CHAVE se deve não apenas por se tratar de um corpo de jornal impresso, mas pela

facilidade de poder usar o mesmo material para algumas comparações também entre as variantes portuguesa e brasileira. Além disso, a possibilidade que o AC/DC oferece de filtrar as buscas por seção do jornal (e, portanto, restringindo as procuras a algumas seções) permite uma comparação mais precisa com a Amazônia. Ou seja, dado que a Amazônia versa sobre assuntos do domínio cultura, e assumindo que os textos têm um viés opinativo, selecionamos, para a comparação, na parte brasileira do CHAVE (jornal Folha de São Paulo), os cadernos "Mais!", "Opinião", "Ilustrada", "Folhateen", "Revista da Folha" e "TV Folha". Na parte portuguesa do CHAVE (jornal PÚBLICO), selecionamos a seção "Cultura". O quadro 2 apresenta detalhadamente os dados do CHAVE, incluindo as informações por seção.

Corpo CHAVE			
Coleções		Público	Folha de São Paulo
Edições		726	730
Documentos		106.821	103.913
Palavras	Total	54.947.072	35.699.765
	Apenas seções "culturais"	5897859	9088172

Quadro 2: Dados do corpo CHAVE

O Museu da Pessoa contém entrevistas realizadas pelo Museu da Pessoa português e pelo Museu da Pessoa brasileiro. O quadro 3 apresenta os dados do corpo Museu da Pessoa por variante.

	Museu da Pessoa (BR)	Museu da Pessoa (PT)
Número de entrevistas	6	107
Palavras	34259	311207

Quadro 3: Dados do corpo Museu da Pessoa

A escolha por esses corpos deve-se às suas características intrínsecas – a ideia é tentar verificar em que medida os textos da Amazônia se aproximam da fala (e, portanto, teriam marcas de oralidade e interação próximas às dos textos do Museu da Pessoa) ou, pelo contrário, estão mais próximos de textos jornalísticos convencionais (e, por isso, o CHAVE). Em um primeiro momento, como indicador de interação entre os interlocutores, buscamos as ocorrências de você (e as variantes voce, voces, vc, vcs), que evidenciariam uma aproximação com o leitor, um suposto diálogo. Antes da comparação propriamente, apresentamos, na tabela 1, a ocorrência de você e das variantes voce e vc na Amazônia, e a distribuição com relação ao total de palavras.

Pronome	Número de ocorrências	Distribuição no corpo Amazônia
Você/você	5434	0.11%
Vocês/vocês	701	0.01%
Vc/vc	110	0.002%
Voce/voce	29	0.0006%
Voces/voces	3	0,0001%
Vcs/vcs	5	0,0001%

Tabela 1: Distribuição de você (e variantes) no corpo Amazônia

Pela tabela 1, é curioso perceber a ocorrência bem maior das formas “você|vocês” com relação a voce|vc|vcs|voces, diferentemente de nossas expectativas com relação à informalidade na escrita do blog/internet. A tabela 2 faz uma comparação entre você/voce/vc/voces/vcs na Amazônia e nos textos do CHAVE e do Museu da Pessoa, agrupados por variante. Como, em Portugal, “você” é usado em um registro formal, para a variante portuguesa consideramos também o tu/te/vós/vos, além dos enclíticos. E, como mencionado, separamos, dentro do CHAVE, os resultados referentes apenas às seções “culturais/opinativas” dos jornais.

Pronome	Amazônia	CHAVE				Museu da Pessoa	
		BR		PT		BR	PT
		Total	“cultural”	Total	“cultural”		
Você (e variações)	0.13%	0.06%	0.1%	0.005%	0.008%	1.2%	0.08%)
Tu/te enclíticos	0.2%	0.0005%	0.01%	0.002%	0.01%	0.06%	0.07%
Vós/vos enclíticos	0.002%	0.0001%	0,0008%	0.0004%	0.002%	0	0.02%

Tabela 2: Distribuição dos pronomes "você"/tu/vós na Amazônia, no CHAVE e no Museu da Pessoa, sobre o total de palavras de cada corpo.

Mesmo que os textos da Amazônia apresentem uma forma de escrita mais convencional do que supúnhamos, ao menos quanto ao item “você” (e variantes), ainda assim o uso do pronome é muito mais frequente na Amazônia que no jornal impresso, o que sugere uma maior interação com o leitor no blog que no jornal. É curioso notar que a frequência de “você” na Amazônia é, inclusive, maior que nas entrevistas do Museu da Pessoa BR. Uma possível explicação para esta alta frequência é a ampla utilização, na Amazônia, do “você” não apenas como

referência direta ao interlocutor – caso das entrevistas – mas também como uma referência genérica, o equivalente ao "se" impessoal. Essa é uma hipótese que pode ser explorada no futuro. Por fim, não é de estranhar a baixa ocorrência de "você" nos textos de Portugal, uma vez que seu uso é restrito a contextos mais formais. No entanto, se olharmos apenas para os resultados das seções de cultura dos jornais impressos, o quadro se altera. A frequência de "você" é praticamente a mesma na parte BR do CHAVE e na Amazônia, o que pode sugerir que esse traço de interação é, na verdade, característico não dos blogs, mas dos textos veiculados pelos blogs. Na variante PT do CHAVE, por outro lado, não observamos o mesmo.

Ainda com relação à interação, buscamos, na Amazônia, a distribuição de alguns marcadores conversacionais: marcadores de hesitação como *ah; eh; humm. Hmm*; marcadores de busca de apoios, como *né?* ; e marcadores de tomada de turno, que indicariam a retomada de um turno da fala, como *Bem, Bom,* (tabela 3). Embora nenhum desses marcadores seja exclusivo da fala, sua ocorrência na escrita costuma ser restrita.

Marcadores conversacionais	Amazônia	CHAVE				Museu da Pessoa	
		BR		PT		BR	PT
		Total	"cultural"	Total	"cultural"		
Bem,	0,003%	0,001%	0.003%	0,0006%	0.001%	0,06%	0
Bom,	0,003%	0,0006%	0.001%	0,0009%	0.001%	0,008%	0,002%
né	0,006%	0,0007%	0,0007%	0,00001%	0	0,8%	0,0003%
Ah/ah/ahh...	0,009%	0,002%	0.005%	0,002%	0.003%	0,1%	0,02%

Tabela 3: Distribuição de marcadores conversacionais nos corpos Amazônia, CHAVE e Museu da Pessoa

Os dados da tabela 3 suportam a ideia de que a escrita nos blogs busca uma reprodução de fala, e uma aproximação com o leitor. A ocorrência dos marcadores, na Amazônia, está no meio do caminho entre o texto de jornal (CHAVE) e a fala (Museu da Pessoa), mas ainda assim se aproxima mais da fala. Um dado curioso é a mesma distribuição, na Amazônia, do "Bem," e "Bom,", que por sua vez têm uma distribuição bem diferente no Museu da Pessoa e, também, por variante. Na fala brasileira, o "Bem," é bem mais frequente como retomada de turno que o "Bom,", embora, na escrita, não haja diferenças. E o "né", como era de se esperar, é praticamente ausente na variante portuguesa. Quando comparamos os dados da Amazônia com os das seções culturais dos jornais impressos, essas diferenças se mantêm, ainda que de maneira mais amena. A exceção está no marcador "Bem,", que tem praticamente a mesma frequência de ocorrência em ambos os corpos.

Quanto aos demais marcadores, a Amazônia continua em uma posição intermediária entre fala e escrita.

Outras pistas exploradas na verificação da expressão de subjetividade foram a presença de verbos e de pronomes na primeira pessoa, e a presença do verbo "achar" verbo típico, na variante brasileira, para a expressão de opinião. A tabela 4 apresenta a comparação entre as ocorrências na Amazônia, no CHAVE e no Museu da Pessoa.

		CHAVE				Museu da Pessoa	
		BR		PT		BR	PT
Tipos de marcadores	Amazônia	Total	"cultural"	Total	"cultural"		
Verbos na primeira pessoa do singular (% sobre o total de verbos)	7%	2,7%	4%	1,8%	3.3%	15%	13%
Verbo achar na 1 pes sing (% sobre o total de verbos na 1ª pes sing)	0,04%	0,02	0.005%	0,003%	0.01	0,18%	0,09%
Meu/minha (% sobre o total de palavras)	0,2%	0,07%	0.1%	0,05%	0.08%	0,5%	0,8%
Nosso/nossa (% sobre o total de palavras)	0.1%	0.06%	0.08%	0.05%	0.06%	0.1%	0.08

Tabela 4: Distribuição de indicadores de subjetividade nos corpos Amazônia, CHAVE e Museu da Pessoa

Novamente, os dados da Amazônia ficam em uma zona intermediária entre o discurso jornalístico e a fala, e isto se aplica a todos os pontos explorados. Mesmo se considerarmos, no texto de jornal, apenas as seções culturais, ainda que as diferenças sejam menores, a Amazônia contém mais verbos na 1ª pessoa, mais ocorrências do verbo "achar" na primeira pessoa e de pronomes de primeira pessoa. No entanto, diferentemente das comparações entre os marcadores discursivos típicos da fala, quando observamos uma ocorrência ligeiramente maior na Amazônia, a expressão de subjetividade é muito mais acentuada na Amazônia - a diferença entre os dados da Amazônia e dos demais corpos é bem marcada, o que corrobora a ideia de textos de blogs como textos altamente pessoais. Embora também seja grande a diferença entre a Amazônia e o Museu da Pessoa, a alta frequência de verbos e pronomes da primeira pessoa neste último está de acordo com as entrevistas do Museu da Pessoa, que versam sempre sobre a história pessoal de vida dos entrevistados. Do lado português, no entanto, chama a atenção a baixa ocorrência do verbo "achar", relativamente á variante brasileira. Uma possível explicação é o uso, nessa variante, do verbo "pensar" no mesmo contexto em que os brasileiros usam o "achar" (no sentido de "crer"; "considerar").

Por fim, e retomando a ideia de caracterização de blogs enquanto gênero, a tabela 5 tenta refletir a opinião que os autores da Amazônia têm sobre o próprio texto que escrevem. Para tanto, foi feita uma busca, na Amazônia, pela estrutura "este/esse *", em que o * corresponde a um caractere coringa cuja classe gramatical é um substantivo[3]. Os resultados (3.801 construções) foram agrupados por lema. A tabela 5 apresenta os lemas mais frequentes em ordem crescente, bem como a frequência associada a cada lema.

Lema		Número de ocorrências
1.	ano	331
2.	tipo	240
3.	**texto**	200 (2.2%)
4.	história	150
5.	**trabalho**	125 (1.4%)
6.	coisa	89
7.	idéia	87
23.	**artigo**	43 (0.4%)
24.	**matéria**	42 (0.4%)
49.	entrevista	28 (0.3%)

Tabela 5: Lemas mais frequentes para a expressão "este/esta /esse/essa *" na Amazônia

Na estrutura buscada, o terceiro lema mais frequente é "texto", o que é bastante sugestivo quanto à forma genérica que os falantes têm de fazer referência à própria produção escrita. Logo em seguida, em 5º, está "trabalho" e, em 23º e 24º, "artigo" e "matéria", respectivamente. Embora "trabalho" possa ser compreendido também como um tipo genérico de texto, não é possível, apenas por esses dados, afirmar que seja este o caso, e o mesmo acontece com "artigo" e "matéria". Uma análise da concordância de cada um desses casos[4] revelou que:

Com relação a "texto", de todas as 200 ocorrências, apenas duas(frases g-h) não eram meta-referência. É interessante perceber que, na frase h, temos também a indicação de como o autor considera o próprio texto: um "relato".

g) id="4339" titulo="o bom e velho jornalismo-esta-morrendo": E quem quiser ler sobre esse processo pode acessar esse texto aqui)

h) id="3935" titulo="o-balanco": Com esse texto da Elisa Lucinda, mesclando a magia da Jornada com a poesia, finalizo esse terceiro e último relato sobre a 12ª Jornada Nacional de Literatura .

Com relação a “artigo” e “matéria”, a situação é bem parecida. De todas as ocorrências, apenas um uso de “artigo” e um de “matéria” não fazem referência ao próprio texto:

i) id="1141" titulo="javier torre diretor argentino parte-1": Eu encontrei na internet esse artigo .

j) id="4638" titulo="como namorar no overmundo parte-1-1": » Muito interessante essa matéria .

Com “trabalho”, a situação é diferente. Das 125 ocorrências, apenas 5 referem-se ao próprio texto, e como é possível observar pelos exemplos, 4 delas elas podem ser enquadradas em um discurso acadêmico (k-n); em 96% das vezes “trabalho” refere-se a algo externo ao texto (frases o-p).

k) id="1841" titulo="sera que ta tudo dominado": Assim, este trabalho compreende que a linguagem se encontra na multiplicidade dos gêneros do discurso, na diversidade de» estilos de linguagem» que expõem aspectos de experiências singulares e coletivas, acontecimentos que devem ser fontes imprescindíveis da analise compreendendo-a como espaço estético de criação cultural .

l) id="4008" titulo="camelodromo da praca xv improviso comunicacão e auto organizacao": Houve também registro, em diário de campo, de conversas informais com camelôs, nas quais estes tratam igualmente dos temas sobre os quais esse trabalho versa .

m) d="3871" titulo="tradicaomodernidade no carimbo de belem": Este trabalho constitui um resumo de minha Dissertação de Mestrado, defendida em 2003 no Instituto de Artes da UNESP/São Paulo

n) id="1860" titulo="sistema de informações financiamento para-musica": Este trabalho está sob uma licença Creative Commons Atribuição -- Uso Não Comercial -- Não a obras derivadas 2.5 Brazil .

o) id="3742" titulo="medea da meia noite ao-amanhecer": E, na verdade, esse trabalho não termina aí, é algo contínuo .

p) id="3963" titulo="lingua e lugar rima região rap e dez de-quadrao": Este trabalho foi mixado no Rio de Janeiro com o Pedro Garcia, baterista dos Seletores de Freqüência, a banda do B Negão .

Um dado revelador é que, dos 8909 resultados, houve apenas uma ocorrência de “blog” e que, ainda assim, faz referência não ao texto propriamente, mas ao blog em que, originalmente, o texto foi publicado.

q) id="4978" titulo="meu nome não e johnny-nem-mauricao": Um dia, recebi um e-mail de um cara dizendo que estava fazendo um livro e que, pesquisando na internet à procura de material, chegou a esse blog .

Vale comentar ainda duas ocorrências de resenha, ambas referindo-se ao próprio texto:

r) id="1968" titulo="a banda tropicalista do duprat-1968": Depois disso, essa resenha perde todo o respeito, eu sei .

s) id="4665" titulo="cravo da terra e o-show-infinito": Para fechar essa resenha remeto ao texto do também ilhéu Chico Saraiva (cujo» CD»» Saraivada» foi lançado em 2007 e cuja resenha fico aqui devendo) que serve de introdução para o site do grupo:»

4. Considerações finais

Apresentamos aqui a Amazônia, uma parte do projeto Floresta Sintá(c)tica composta por textos de um blog. Como parte importante da documentação da Amazônia, temos o desafio de enquadrá-la em termos de gênero textual.

Em uma análise preliminar podemos caracterizar os textos da Amazônia/Overblog em entrevistas, resenhas, reportagens, artigos e narrações/relatos de eventos. Por isso, em termos de gênero, usar blog pode não fazer sentido, visto que todo o conteúdo parece poder ser descrito como em um gênero tradicional, o que não chega a ser novidade. Como reconhece Marcuschi (2005), boa parte dos chamados gêneros emergentes possui similares em outros ambientes. No entanto, há nítidas influências do suporte no texto – em termos não apenas de informalidade, mas principalmente em termos de aproximação com a fala, como mostra a breve comparação entre os textos da Amazônia e textos de jornal impresso, por um lado, e de entrevistas, por outro. Ainda assim, é interessante notar que, quando comparamos os dados da Amazônia com apenas as seções "culturais" do jornal impresso, em alguns casos as diferenças ficam menos evidentes.

Assim, por um lado, se caracterizamos a Amazônia como blog, deixamos de capturar dados relevantes, como a presença de entrevistas, resenhas, artigos etc. Se desprezamos o rótulo blog, deixamos de capturar regularidades relevantes para a caracterização dos textos ali inseridos, como a subjetividade e a interação.

Especificamente, se apresentamos a Amazônia é um corpo composto por blogs, não somos capazes de oferecer uma informação mais precisa sobre o seu conteúdo – tanto em termos de forma, quanto de função. Dizemos apenas que são textos que contêm um forte viés subjetivo, de expressão de opinião. Mas, se considerarmos que blog também pode veicular entrevista, ou uma narrativa, uma receita culinária, a descrição é pobre, pois perde as nuances que a CMC carrega.

Some-se à dificuldade inerente à caracterização dos gêneros relacionados à "revolução da Internet" (para usar os termos de Crystal (2004) a efemeridade e rapidez que caracterizam os seus "objetos", o que torna qualquer tentativa de caracterização, descrição, ainda mais parcial e contingente do que já são, normalmente, as tentativas de descrição e caracterização de qualquer objeto. Prova disso é que, se entre 2004-2007 são comuns às menções à revolução dos blogs e à blogosfera, agora já se fala em tuitosfera, numa junção de blog e twitter, em que o segundo aparece como o "objeto revolucionário" da vez, e começam a surgir trabalhos na área de PLN que buscam caracterizar a linguagem usada no twitter (Fernandes et al., 2010)

Gêneros textuais são fenômenos sociais e históricos. Quanto aos blogs, especificamente, questionar o seu enquadramento em termos de um novo gênero não significa questionar a influência do suporte nos gêneros textuais. Enquanto objetos históricos, é natural que os gêneros sofram mudanças/ ampliação – o que se observa, por exemplo, nos bate-papos e páginas pessoais. No entanto, nem todo

texto apoiado em um novo suporte talvez mereça ser designado um novo gênero. Leis já foram escritas em paredes, em tábuas, pergaminhos, papel e editores de texto, e nem por isso deixam de ser leis.

Com o reconhecimento de que não se trata de um novo gênero, mas de gêneros tradicionais com diferenças decorrentes das características do suporte, abre-se espaço para uma descrição comparativa assentada nas diferenças que podem, inclusive, ser tão gritantes que forcem de fato a inclusão de blogs como um novo gênero. Mas tal só poderá ser feito se partirmos do muito de comum – em termos de movimento de texto, de função, de objetivos – que existe entre os diversos textos abarcados/veiculados pelo blog. A disponibilização de um corpo com tais características é o primeiro passo em direção a uma descrição mais sistemática desses “novos” textos.

Agradecimentos

A Amazônia e o projeto Floresta Sintá(c)tica são desenvolvidos no âmbito da Linguateca, co-financiada pelo governo português, pela União Européia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN.

Referências

AFONSO, Susana, BICK, Eckhard, HABER, Renato & SANTOS, Diana. "Floresta sintá(c)tica: um treebank para o português". In GONÇALVES, Anabela & CORREIA, Clara Nunes (eds.), *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística* (APL 2001) (Lisboa, 2-4 de Outubro de 2001), Lisboa, Portugal: APL, pp. 533-545.

ANDREEVSKAIA, Alina, BERGLER, Sabine & URSEANU, Monica. "All blogs are not made equal: Exploring genre differences in sentiment tagging of blogs". In *International Conference on Weblogs and Social Media*. Hilton Seattle Downtown, Seattle, Washington, U.S.A. 2007

BERBER SARDINHA, Tony. *Análise de Gênero e Linguística de Corpus: Identificação das unidades internas do gênero por meio da padronização lexical.* DIRECT Papers 51, ISSN 1413-442x 2003. LAEL, 2003.

BICK, Eckhard, SANTOS, Diana, AFONSO, Susana & MARCHI, Rachel. "Floresta Sintá(c)tica: Ficção ou realidade?". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 291-300.

BICK, Eckhard. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Dr.phil. thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press. Novembro de 2000.

BLOOD, Rebecca. "Weblogs: A History and Perspective" Rebecca's Pocket, Sept. 7, 2000; disponível em www.rebeccablood.net/essays/weblog_history.html.

CIAPUSCIO, Guiomar Helena. "Famílias de gêneros e novas formas comunicativas para a ciência". *Calidoscópio*. Vol. 7, n. 3, p. 243-252, set/dez 2009.

COSTA, Luís, SANTOS, Diana & ROCHA, Paulo Alexandre. "Estudando o português tal como é usado: o serviço AC/DC". In *The 7th Brazilian Symposium in Information and Human Language Technology* (STIL 2009) (São Carlos, SP, Brasil, 8-11 de setembro 2009).

CRYSTAL, David. *A Revolução da Linguagem*. Rio de Janeiro: Jorge Zahar, 2005.

FREITAS, Cláudia & AFONSO, Susana. "Bíblia Florestal: Um manual lingüístico da Floresta Sintá(c)tica". 29 de Setembro de 2008. http://www.linguateca.pt/Floresta/BibliaFlorestal/completa.html

FREITAS, Cláudia, ROCHA Paulo Alexandre & BICK, Eckhard. "Um mundo novo na Floresta Sintá(c)tica - o treebank para Português". *Calidoscópio*. Vol. 6, n.3 pp. 142-148. 2008.

FREITAS, Larissa Astrogildo de, FERNANDES, Angélica Alves & CORRÊA, Ulisses Brisolara. "Minerando Tweets". Pôster apresentado no ELC 2010.

HERRING, Susan, SCHEIDT, Louis A. & BONUS, Sabrina. "Weblogs as a bridging genre". *Information Technology & People*; vol 18, n. 2, 2005. pp 142-171.

LESHED, Gilly & KAYE, Joseph. Understanding How Bloggers Feel: Recognizing Affect in Blog Posts. In *Proceedings of CHI-2006*, Montreal, Canada.

MARCUSCHI, Luiz Antônio. "Gêneros textuais emergentes no contexto da tecnologia digital". In Luiz Antônio Marcuschi & Antonio Carlos Xavier (Eds.), *Hipertexto e Gêneros Digitais*. Rio de Janeiro: Editora Lucerna, 2005.

MAURANEN, Anna. "Another look at genre: Corpus Linguistics vs Genre Analysis". *Studia Anglica Posnaniensia*, 32, 303-315, 1998.

MILLER, Carolyn R. & SHEPERD, Dawn. "Blogging as social action: A genre analysis of the weblog". In Laura Gurak, Smiljana Antonijevic, Laurie Johnson, Clancy Ratliff & Jessica Reyman (Eds.), I*nto the Blogosphere. Rhetoric, Community, and Culture of Weblogs*. 2004.

MILLER, Carolyn R. "Genre as a social action". *Quarterly Journal of Speech*, Vol 70, pp.151-167. 1984.

SANTOS, Diana. "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, 2008, pp.41-66.

SWALES, John. *Genre Analysis - English in Academic and Research Settings*. Cambridge: Cambridge University Press, 1990.

[1] Disponível em http://www.linguateca.pt/ACDC/

[2]O projeto AC/DC está acessível de www.linguateca.pt/ACDC

[3] No AC/DC, a expressão de busca utilizada foi [word="[Ee]s[st][ea]"] @[pos="N.*" & id="AMAZ.*"], e a forma de exibição dos resultados foi "distribuição de lema".

[4] Para a análise da concordância de cada uma das palavras-alvo, utilizamos a expressões de busca [word="[Ee]s[st][ea]"] @[lema="trabalho" & id="AMAZ.*"], e a forma de exibição dos resultados foi "concordância"