Informática e literatura: revelando identidades textuais

Tania M. G. Shepherd

Situando o computador na análise textual

Sem querer reduzir os avanços na área computador-texto a uma mera listagem, é necessário dizer que as décadas de oitenta e noventa concentraram-se em três grandes blocos de atividades no que tange aos estudos textuais auxiliados por computador. O primeiro deles enfocou os diferentes métodos de compilação de *corpora* propriamente ditos e modos de anotação (codificação, etiquetação e segmentação). Os exemplos de pesquisa nessa área são inúmeros, mas podemos mencionar o trabalho de Douglas Biber (1989, 1990 e 1994) sobre tipologia de textos de língua inglesa, incluindo ai também o texto de ficção.O trabalho de Biber resultou em uma nova gramática da língua inglesa, a primeira a categorizar a fala e a escrita como entidades distintas, empiricamente investigadas. Desde o inicio da compilação desses *corpora*, inúmeros outros foram compilados já agora com anotações que permitiam a garimpagem de elementos sintáticos e discursivos (Carter e McCarthy, 1995). As anotações de *corpora*, em grande parte de natureza lingüística, ainda que laboriosas, tornavam a caracterização daquilo que se investigava muito mais rápida e quantitativamente representativa (Leech 1987, 1993). Um dos trabalhos oriundos dessa tradição é Short e Semino (no prelo) que usa a metodologia da anotação de *corpora* formados de textos em inglês de ficção e não ficção produzidos no fim do século XX, para investigar os diferentes modos de apresentação da fala e do pensamento nesses textos, configurando o que podemos chamar de um trabalho de fôlego sobre um novo ramo da investigação de *corpora* chamado estilística de corpus.

A segunda área de enfoque desses estudos consistiu nos vários modos de exploração desses *corpora* ou de arquivos textuais. Com o desenvolvimento das ferramentas computacionais para análise textual e seu lançamento no mercado para uso individual e institucional, tornou-se mais fácil garimpar textos diversos. O maior avanço nesta área foi a disponibilização dessas ferramentas na própria Internet e sua aplicação à Internet, fazendo com que essa se tornasse um *corpus* ilimitado ou como Renouf (2003) rotulou, um *corpus sans frontiers*.

A terceira área de concentração surgiu no final dos anos oitenta, quando o modo de marcação dos *corpora*, de base sintática, passou a ser preterido por outras marcações como as 'colocações', grupos de palavras ordenadas não por critérios sintáticos ou mesmo por marcações semânticas e prosódicas. Começam nessa época os questionamentos sobre o que seriam as unidades mínimas de análise com Sinclair (1996) o qual questiona a sentença e seus componentes como unidades analíticas adequadas. (cf. Berber-Sardinha: 2000)

Quanto aos estudos literários auxiliados por métodos computacionais , esses datam de pelo menos a década de quarenta. No princípio, o trabalho de compilação de concordâncias era feito manualmente. Subseqüentemente era feito com perfuração de cartões para serem lidos por máquinas separadoras e, finalmente por computadores de grande porte (*mainframe*). O auxílio do computador nesses estudos foi uma tentativa, a principio de substituir a abordagem pessoal do crítico pelo potencial dos resultados numéricos e estatísticos, abordagem esta cercada da objetividade nem sempre bem acolhida pelas pessoas da área. Finalmente, em 1986, entretanto, faz-se uma marca referencial nos estudos literários assistidos por computador. Nesse ano começa a ser editado o periódico Literary and Linguistic Computing, consagrando de vez por todas a interface dessas duas áreas.

1.1. Tipos de estudo

Os estudos auxiliados por computador fazem uma importante distinção entre o que se convencionou chamar *corpus-based*, estudos que partem de modelos pré-estabelecidos baseados em corpus (Clear et al. 1996). e aqueles chamados de *corpus-driven*, estudos que partem do corpus sem modelos ou *insights* prévios, para do corpus extrair possíveis padrões de uso ou observar a ausência deles. Dentro das investigações de textos literários, Hunston (2000: 128) admite que o estudo de *corpora* digitalizados tem um lugar ambíguo já que há uma séria resistência por parte dos acadêmicos quanto à consubstanciação de suas interpretações individuais através do que chamam de evidências textuais. Entretanto, como Leech e Short (1981) já haviam apontado quando discutiam o estudo estilístico, há necessidade de algum tipo de evidência para embasar aquilo de subjetivo que o analista pode oferecer em uma análise. Mesmo que essa evidência venha através da comparação dos *corpor*a literários com aqueles de uso cotidiano, como já sugerido por Louwn (1997) e Barnbrook (1998), citados por Hunston (op.cit.)

Com relação especifica à pesquisa estilística de *corpora* digitalizados, formados de textos de literatura[1] as perguntas de pesquisa tendem a abordar os seguintes pontos:

a) estilos individuais de autores
b) estabelecimento de autoria
c) características de uma obra individual.
d) características de um gênero específico
e) características de um período literário: estilística histórica

Hunston (op.cit: 42-43) sumariza bem essas múltiplas aplicações ao dizer que o computador seleciona aquilo que é mais freqüente no texto, o que forma padrões observáveis, corroborando, dessa forma, intuições nem sempre facilmente transparentes para outros que não o analista. Ao mesmo tempo, o computador também nos ajuda a observar o raro, o marginal, o atípico e o caso único. Ao se compararem dois *corpora,* mostra-se o que é característico

de cada um deles. Entretanto, apesar de tão útil essa interface computador-

literatura, Hunston (op.cit: 130) menciona a dificuldade de se ter um acadêmico da área literária com a disposição e o *expertise* de automatizar seus processos de marcação textual[2]. As razões que a autora atribui são ou porque a percepção do texto literário é que ele resiste uma segmentação ou compartimentalização ou porque as categorias analíticas para a exploração do texto literário digitalizado não são bem escolhidas *a priori.*

III. Novos modos de olhar o texto literário

Com estas restrições citadas por Hun ston em mente, nesta seção descreverei dois *insights* sobre textos literários obtidos através do computador. O primeiro *insight,* mais simples, gira em torno da identificação de uma pequena mas marcante característica genérica da narrativa policial: Na análise manual das escolhas lexicais de um conto pós-moderno, do autor inglês John Fowles, conto esse pretensamente policial, Shepherd (2001) observou-uma determinada preferência por certa organização lexical que possivelmente ecoaria as histórias detetivescas tradicionais. Uma varredura dos parágrafos iniciais de tão somente um dos contos de Conan Doyle[3] mostrou uma preponderância de colocações semelhantes, ou seja o uso de itens pertencentes ao campo lexical de ´estranho´ (em língua inglesa: *odd, singular, unusual, bizarre*), bem como a presença de superlativos atrelados a estes itens lexicais, exemplificados a seguir:

a) "Now, Mr. Jabez Wilson here .... and to begin a narrative which promises to be *one of the most singular* which I have listened to for some time."

b) "As far as I have heard, it is impossible for me to say whether the present case is an instance of crime or not, but the course of events is certainly *among the most singular* that I have ever listened to."

c) "I cannot see that there is anything very funny," cried our client, flushing up to the roots of his flaming head. "If you can do nothing better than laugh at me, I can go elsewhere." "No, no," cried Holmes, shoving him back into the chair from which he had half risen. "I really wouldn't miss your case for the world. *It is most refreshingly unusual.*"

d) "As a rule", said Holmes, "*the more bizarre a thing is the less mysterious it proves to be*. It is your commonplace, featureless crimes which are really puzzling, just as a commonplace face is the most difficult to identify. But I must be prompt over this matter." "What are you going to do, then?" I asked.

Uma investigação mais aprofundada nos demais contos de Doyle com o auxilio do computador também revelou ser essa combinação de nuances do significado de estranho marcadas por graus comparativo e superlativo caracterizadoras de vários outros contos de Doyle. A conclusão inicial tirada foi que Fowles se utilizaria de formas con sagradas do conto policial– e marcadas lexicalmente- para descontruir a sua história de detetive. Este insight, depois aplicado a outro conto policial que deliberadamente boicota as fórmulas con sagradas, *A morte e a bússola* de Jorge Luis Borges., sugere que a apropriação

dessas fórmulas, que vão muito além do empréstimo meramente temático, pode ser verificada computacionalmente. Obviamente, para uma afirmativa de mais vulto, seria necessário organizar um corpus tão somente de histórias policiais e organizar um estudo exaustivo destas marcas genéricas.

Nosso segundo comentário tem a ver com os resultados da análise assistida por computador de *O Papagaio de Flaubert.* Antes porém são necessárias algumas palavras sobre o romance, que é considerado um dos melhores produzidos por Julian Barnes, escritor inglês contemporâneo, dado a experimentações narrativas. Publicado em 1984, *O Papagaio de Flaubert* foi finalista do Booker Prize de 1984 e ganhou o aclamado Prix Medicis na França. Em comum com muitos romances rotulados de pós-modernos, ele mostra descontinuidade temporal que, segundo Lodge (1977) é a chancela desse tipo de narrativa. *O Papagaio de Flaubert* porém opera descontinuidade em múltiplos níveis, além de uma simples continuidade temporal.

O primeiro e último capítulos (1 e 15) abrem e fecham a história da busca do narrador pela identidade de um certo papagaio empalhado que serviu de modelo para o papagaio de *Un coeur simple*, um dos contos de Flaubert. . Assim, o capítulo 1 termina com o narrador enumerando as várias cartas enviadas a especialistas para traçar a origem do referido papagaio:

> *After I got home the duplicate parrots continued to flutter in my mind: one of them amiable and straightforward, the other cocky and interrogatory. I wrote letters to various academics who might know if either had been properly authenticated... I hoped to get my replies quite soon. (F.P. p. 22).*

No capitulo 15, ou treze capítulos depois do que podemos rotular de uma longa digressão, o narrador retoma e deixa sem resolução o enigma inicial da narrativa, dizendo:

> *"And the parrot? Well it took me almost two years to solve the Case of the Stuffed Parrot. The letters I had written after first returning to Rouen produced nothing useful; some of them weren't even answered." (F.P. p. 180)*

Histórias com enigmas não resolvidos não são novidade. Entretanto, entre os capítulos 1 e 15 do referido romance, são justapostos, de maneira quase aleatória[4], treze outros capítulos que deixam de lado o tema do mistério do papagaio. Cada um deles é escrito em simulacros de gêneros textuais completamente distintos. Por exemplo, o capitulo 2 pede emprestado o seu formato às cronologias. O capitulo 4 é um catálogo de nomes de animais. O capitulo 6 é um artigo de critica literária. Há também um capitulo que se assemelha a um dicionário e outro que é um exame final sobre literatura. Além dessa organização pouco convencional, individualizada em cada um dos gêneros adotados para os capítulos, o que une os capítulos de *O Papagaio de Flaubert* não mostra as condições necessárias e suficientes que compõem o discurso narrativo, quais sejam, um mínimo de temporalidade e causalidade.

O que temos, portanto, são treze gêneros não narrativos sobre temas distintos formando o recheio de dois capítulos narrativos. Entretanto, a despeito de toda essa experimentação, o livro foi escrito, editado, impresso, vendido e premiado como um romance, um processo de produção que o torna *hiper-protegido* (Pratt ,1977) contra qualquer tentativa do leitor de tratá-lo de forma diversa.

Como para nós não deve bastar dizer que um livro é um romance para o consideramos como tal, resolvemos também embarcar numa busca de identidade para *O Papagaio de Flaubert*.

O primeiro passo da pesquisa foi digitalizar o livro, processo longo e laborioso que envolveu digitação e escaneamento – duas estratégias constantes de toda a transcrição de textos que não são originariamente concebidos em forma digital. Uma vez em formato de símbolos que poderiam ser lidos por máquina (ASCII), todas as sentenças do livro (qualquer espaço entre dois pontos finais) foram numeradas através de um programa computacional desenhado originariamente para fazer sumarização automática de textos não narrativos. O objetivo maior dessa leitura através de um computador não foi para testar o *software* com um texto tão longo quanto um romance, mas sim para poder iluminar a organização de um texto que à época foi reconhecidamente difícil de classificar e, por falta de titulo melhor, foi rotulado como "biografia pós-moderna" ou seja uma biografia com muitos traços não narrativos.

Sobre o programa propriamente dito, comissionado pela British Telecom à Universidade de Birmingham, Inglaterra nos anos noventa, ele tinha como objetivo original a extração automática de sumários de textos. As premissas por trás do programa eram provenientes do trabalho sobre coesão textual e padrões lexicais desenhado por Hoey (1991) segundo o qual:

a) as frases que têm itens lexicais em comum também têm tópicos em comum.

b) em textos não narrativos quaisquer frases que tenham um mínimo de três repetições lexicais em comum formam um elo e frases com três elos estão obviamente proximamente relacionadas.

c) frases com três elos em qualquer parte de um mesmo texto e em qualquer distância umas das outras fazem sentido e são coerentes quando justapostas.

d) ao serem justapostas todas as frases com três elos de um texto não narrativo, pode-se formar um sumário automático do texto em questão.

Um dos problemas analíticos do livro de Barnes con sistia em como um amontoado de capítulos que poderiam ser lidos aleatoriamente poderia conter algum tipo de coesão. Uma outra pergunta era que capítulos uniam-se para permitir ao leitor um mínimo de processamento do texto. No caso especifico de *O Papagaio de Flaubert* nosso objetivo era primeiro verificar se nossa intuição sobre os capítulos 1 e 15 se confirmaria, isto é se fossem realmente narrativos, não produziriam nem sumário. Em contrapartida, os demais capítulos obviamente não narrativos deveriam produzir algum tipo de sumário através de elos coesivos, cuja natureza mereceria ser investigada. Em outras palavras, os capítulos narrativos, quando justapostos produziriam sumários incoerentes e o oposto aconteceria com os capítulos não narrativos.

Não vamos aqui entrar nos detalhes analíticos dos elos textuais formados após a aplicação do programa[5]. O que podemos dizer, em síntese é que os elos estabelecidos por determinados capítulos formavam sumários coerentes e que pode verificar-se a densidade dessa coerência, ou seja que capítulos continham mais densidade de ligação que outros. Verificou-se também que o capitulo 1, o início do sintagma narrativo da história do papagaio de Flaubert continha elos com cada um dos demais capítulos, e pouquíssimos elos com o capitulo final. Um desses pouquíssimos elos pode ser exemplificado nas frases 140 - 3993

> *[140] Psittacus, ran the inscription on the end of its perch; 'Parrot borrowed by G. Flaubert from the Museum of Rouen and placed on his work-table during the writing of Un coeur simple, where it is called Loulou, the parrot of Félicité, the principal character in the tale'.*
> *[3973] I lay in bed on my back with the lights out and thought about Flaubert's parrot: to Félicité it was a grotesque but logical version of the Holy Ghost; to me a fluttering, elusive emblem of the writer's voice.*

Esses elos entre o primeiro e o ultimo capitulo, pouquíssimos em número, sugerem que a consistência coesiva entre esses dois blocos é de natureza narrativa, já que o programa não gera pares coerentes em discurso narrativo. Tal achado foi diferente nos demais capítulos não narrativos, que não só formavam elos entre si como também através do romance como um todo. O mais interessante é que, conforme descrito por Hoey (1991)há pouca probabilidade de geração de elos quando os segmentos estão separados por muito texto, o que ficou caracterizado como possível após a análise dos elos de *O Papagaio de Flaubert*[6].

Conclusões

Com a inversão da temática do seminário neste pequeno trabalho, não foi minha proposta delinear tão somente duas das múltiplas aplicações do uso do computador na garimpagem do texto literário. O computador não fará nada, a não ser que o analista saiba fazer a ele as perguntas certas. No atual estado de conhecimento, ele poderá ser um valioso instrumento na análise da organização, do estilo, do vocabulário e conteúdo dos textos em geral e dos textos de ficção em particular, intra-texto e inter-texto, se e somente se as perguntas certas lhe forem feitas. Parafraseando Smedt (2002:92), trabalhar com representações formais ou com quantificação de problemas acadêmicos implica a geração de novos conhecimentos através de novas atitudes frente ao objeto analítico. A metáfora do garimpo por nós usada é portanto adequada. O computador é somente a bateia. O produto do garimpo somente advém quando o analista sabe como e onde garimpar. De qualquer forma, e como a presente mesa corrobora, não viemos aqui discutir os múltiplos usos da máquina, mas sim alguns novos modos de pensar o texto, novos pontos de partida para uma análise textual.

Referências bibliográficas:

BERBER-SARDINHA, T: (2000) Lingüística de Corpus: histórico e problemática. *D.E.L.T.A*. Vol. 16, nr2, 323-367.

BIBER, D. (1989) *A Typology of English Texts*. Linguistics 27:3-43.

BIBER, D. (1990) *Methodological Issues Regarding corpus-based Analyses of Linguistic variation*. Literary and Linguistic Computing 5:4:257-269.

BIBER, D., S. Conrad & R. Reppen (1994) *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: CUP.

CARTER, R. & McCarthy, M (1997) *Exploring Spoken English.* Cambridge: CUP.
CLEAR, J. et al. (1996) "COBUILD, The State of the Art". *International Journal of Corpus Linguistics* 1:2:303-314.

GRANGER, Sylviane and Stephanie Petch-Tyson, ed. (2003) *Extending the Scope of Corpus-Based Research: New Applications, New Challenges*. Rodopi.
HOEY, M (1991) *Patterns of Lexis in Text*. Oxford: OUP.

HUNSTON, S (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

LEECH, G (1987) "General Introduction". In R. Garside et al. (orgs.), *The Computational Analysis of English - a corpus-based approach*. Longman: London. pp. 1-15.

LEECH, G (1993) "Corpus Annotation Schemes". *Literary and Linguistic Computing* 8:4: 276-281.

LODGE, D (1977) *The Modes of Modern Writing: Metaphor, Metonymy and the Typology of Modern Literature*. London: Edward Arnold.

PRATT, M.L. (1977) *Toward a Speech Act Theory of Literary Discourse.* Bloomington: Indiana University Press

QUIRK, R. (1992) On corpus principles and design. In J Svartvik (org.) *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82. Stockholm, 4-8 August 1991. Mouton de Gruyter, Berlin.pp. 457- 469.
RENOUF, A (2003) Webcorp: the internet as corpora. In Granger, Sylviane and Stephanie Petch-Tyson (orgs). *Extending the Scope of Corpus-Based Research: New Applications, New Challenges*. Rodopi.

SEMINO, E & Short, M (no prelo) *Corpus Stylistics*. *Advances in Corpus Linguistics*. London :Routledge

SINCLAIR, J (1996) The Search for Units of Meaning. *Textus*. Vol 9:75-105.
SINCLAIR, J (ed.) (1987) *Looking Up: An Account of the Cobuild Project in Lexical Computing.* Collins, London.

SHEPHERD, T.M.G. (1997) "Towards a description of atypical narratives: a study of the underlying organization of Flaubert´s parrot". *Language and Discourse*. Vol 5. pp 7 1-97.

SMEDT, K (2002) "Some Reflections on Studies in Humanities Computing". *Literary and Linguistic Computing*, vol 17, n 1, pp. 89-101

SVARTVIK, J. (1992) Corpus linguistics comes of age. In J Svartvik (ed) *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82. Stockholm, 4-8 August 1991. Mouton de Gruyter, Berlin. pp. 7-13.

ZYNGIER, S & Shepherd, T. M G (2003) "What is literature, really: a corpus-driven study of students' statements". *Styl*e, vol 37, n 1, Spring 2003. pp 15-27.

Notas:

[1] Há também o que se convencionou chamar de estilística forense.

[2] Uma exceção aqui é o trabalho de David Myal, da Universidade de Alberta, cujo trabalho computadorizado sobre os românticos da literatura inglesa foi apresentado na UERJ/UFF em 2001 e agora faz parte de um Cd-rom.

[3] O conto em questão é *The Red-Headed League* (1891) e as 'concordâncias' listadas foram obtidas *online* através de comandos dados ao arquivo de contos de autoria de Conan Doyle no site httpi/www.sherlockian.net/canon/index.html.

[4] O termo aleatório aqui é usado propositalmente. Com exceção do capitulo que é um simulacro de uma prova para ver se o leitor apreendeu os conceitos passados no livro, a ordem de leitura dos capítulos é irrelevante.

[5] Essa análise é parte de minha pesquisa de doutorado intitulada A computational analsysis of repeated elements in fringe narratives, submetida à Universidade de Birmingham em 1993 e publicada em Shepherd (1997)

[6] Tal *insight* permitiu que se olhasse, em pesquisas posteriores, a natureza de elos na linguagem jornalística entre periódicos publicados com muito espaço de tempo e entre periódicos distintos. Este foi um passo positivo numa possível identificação de marcas de intertextualidade.

Tania M G Shepherd é PhD pela Birmingham University, Inglaterra. Professora-adjunta na UERJ em língua inglesa e linguística e pesquisadora UERJ/Faperj (2002-2005) com projeto sobre análise textual e computação.