FROMM, Guilherme. Lingüística Computacional: uma intersecção de áreas. In: **Revista Factus**. nº 5. Taboão da Serra: FTS, 2006. p. 135/140.

Lingüística computacional: uma intersecção de áreas

Guilherme Fromm

Resumo: o presente artigo visa apresentar algumas características de uma nova área de pesquisa, na confluência entre Lingüística e Computação, a Lingüística Computacional. Apresentaremos, também, algumas de suas aplicações, já bastante utilizadas ou ainda incipientes.
Palavras-chave: Lingüística, Computação, PLN, Lingüística de *Corpus*

Abstract: this article aims presenting some characteristics of a new research area, in the intersection of Linguistics and Information Technology, the Computational Linguistics. We also show some of its new programs, in usage or still in development.
Keywords: Linguistics, Information Technology, NLP, Corpus Linguistics

Computação: substantivo feminino, ação ou efeito de computar 1 cômputo, cálculo, contagem; operação matemática ou lógica realizada por regras práticas preestabelecidas Ex.: <c. de um prazo> <c. de uma dívida>; 2 Rubrica: informática. m.q. processamento de dados; 3 Derivação: por metonímia. ação ou atividade exercida por meio de computadores eletrônicos

Lingüística: substantivo feminino, Rubrica: lingüística. ciência que tem por objeto: (1) a linguagem humana em seus aspectos fonético, morfológico, sintático, semântico, social e psicológico; (2) as línguas consideradas como estrutura; (3) origem, desenvolvimento e evolução das línguas; (4) as divisões das línguas em grupos, por tipo de estrutura ou em famílias, consoante o critério seja tipológico ou genético

Os estudos sobre língua já vêm sendo praticado há milênios e temos uma bibliografia impressionante sobre o assunto. Os estudos sobre computação praticamente surgiram no século passado e igualmente apresentam um grande número de publicações. Embora, à primeira vista, se configurem como ciências em campos díspares (humanas e exatas), desde o final do século XX e, especialmente a partir do século XXI, têm trabalhado juntas para o aprimoramento de ambas.

Durante muito tempo a Lingüística (ou a gramática, quando nos referimos aos estudos anteriores ao séc. XX) procurou, através de conceitos gregos de analogia e anomalia, desenvolver regras (e exceções) para o bom aprendizado das diversas línguas. Uma das problemáticas para esses estudos era exemplificar essas regras e exceções a questão da exemplificação. Muito do que era criado não passava de invenções dos lingüistas que queriam provar algum ponto de vista. O mesmo se dava na criação de

obras lexicográficas/terminográficas (dicionários, vocabulários e glossários), onde os autores, além de darem uma definição, na maioria das vezes, não muito científica, também criavam exemplos por vezes absurdos.

Para evitarmos esses abusos e abrangermos novas áreas de estudos lingüísticos que possam trabalhar fatos mais concretos, surge essa intersecção entre as áreas de Lingüística e Computação. Uma das primeiras novas áreas a brotar dessa união foi a moderna Lingüística de *Corpus*. Embora não seja uma ciência totalmente nova[1], os estudos sobre *corpora*, no final do século XX e no século XXI, são baseados em análises computacionais.

A Lingüística do *Corpus* propõe novas abordagens de como levantar um *corpus* de acordo com o estudo desejado que se fará a partir dele. Não basta apenas copiarmos e colarmos textos de várias fontes, por exemplo, e depois estudá-los. Necessita-se um equilíbrio entre as áreas estudadas, a quantidade de textos, as línguas abordadas (se não for monolíngüe), etc. Estabelecidos os parâmetros para a construção do mesmo, começa a busca por textos que o alimentem.

O processo mais comum, hoje em dia, é fazer um levantamento através da Internet. Embora a rede trabalhe com excessos (de informação e desorganização), ainda é mais fácil filtrar dados recebidos por meio eletrônico do que transformar dados impressos ou orais nos mesmos. *Corpora* criados a partir de uma base totalmente eletrônica podem ter centenas de milhões de palavras em seu corpo. Teoricamente, quanto mais palavras, ou seja, quanto maior for o *corpus* de estudo (em qualquer área), mais precisos serão os resultados.

A área de computação já trabalha nesse nível de coleta. Existem, além dos instrumentos de busca tradicionais da Internet (como o Google), outros desenvolvidos especialmente para esse fim como o Webcorp (www.webcorp.org.uk), que nos ajudam na seleção dos textos.

Com o *corpus* pronto, buscamos programas de análise de acordo com as nossas necessidades. Se quisermos, por exemplo, uma análise de colocações, co-ocorrências, padrões sintático-semânticos, etc., podemos usar a ferramenta Wordsmith Tools (www.lexically.net). Enquanto o WordSmith trabalha com uma grande quantidade de textos e é indicado para levantamento de termos específicos para vocabulários técnicos

[1] Um *corpus* é um conjunto de textos (escritos e/ou orais) selecionados, dentro de uma área específica, para serem analisados por um lingüista. Daí podem surgir pesquisas em diversas subáreas da Lingüística: lexicografia, terminologia, análise da conversação, análise do discurso, ensino e aprendizagem, tradução.

e lexemas para dicionários (usando aqui as concepções propostas por BARBOSA, 2001), o STABLEX (www.pirus.com.br) é mais recomendado para análises que envolvam discurso: nele podemos identificar, através de poucos textos de um autor, por exemplo, traços de estilo e preferência no uso de palavras.

A Lingüística Computacional facilita não apenas o planejamento, construção e análise de *corpora*, mas também ajuda a vida do consulente de uma obra lexicográfica, por exemplo. Novos dicionários de língua, assim como vocabulários de áreas técnicas, são disponibilizados em versão eletrônica, facilitando bastante a busca de verbetes na macroestrutura da obra. Uma das características mais interessantes dessas obras é a possibilidade de qualquer palavra na microestrutura do verbete servir de *hyperlink* para sua explicação: quando, dentro de uma definição, por exemplo, não entendemos uma palavra, basta dar um duplo clique nela para abrir um novo verbete. As novas possibilidades da informática aumentam bastante a velocidade de consulta, especialmente por parte dos maiores usuários de dicionários, os tradutores.

Para os profissionais da computação, os *corpora* também servem como fonte de treinamento de programas específicos para análises lexicais, sintáticas e semânticas. Um exemplo bastante conhecido de programa, ou parte de programa, criado por desenvolvedores, é o corretor ortográfico dos processadores de texto. Na mesma linha, porém ainda sem tanta precisão no português, caminham os corretores gramaticais.

Para criar e treinar esses programas na análise de textos, são inseridos, nos textos do *corpora*, cabeçalhos (que também ajudam os lingüistas) e etiquetas que indiquem, por exemplo, a classe gramatical de cada palavra. Isso gera também mais um grande desafio para os profissionais: a criação de etiquetadores automáticos. Essas ferramentas ainda estão longe da perfeição: algumas, como etiquetadores de classe gramatical, já conseguem um índice de acerto na faixa de 90%. Embora seja uma grande evolução em relação ao processamento manual por parte de especialistas, esse número deve, teoricamente, ser de 100%; qualquer número inferior significa uma revisão, um novo trabalho por parte dos lingüistas, talvez ainda mais exaustivo.

Pesquisadores da área de computação também buscam o sonho de muitas firmas e pessoas ao redor do mundo: tradutores automáticos que realmente funcionem[2]. Embora já tenhamos vários tradutores (como o Delta Translator) e alguns programas de

[2] Essa área é designada como PLN (Processamento de Linguagem Natural) no universo da computação.

memória de tradução[3] (como o Trados), eles ainda estão longe de serem totalmente automatizados. As traduções geradas por esses programas requerem sempre uma revisão; muitos deles já trabalham com vocabulários de especialidade, que podem ser baixados da Internet, mas mesmo assim cometem erros. Os programas de memória de tradução requerem um exaustivo pré-trabalho de elaboração de glossários que serão incorporados à determinada tradução (mais com o intuito de padronizá-la e barateá-la do que automatizá-la completamente). Embora projetos nessas áreas soem como maravilhas, já que dispensariam totalmente o profissional de tradução, a realidade, no momento, nos mostra que ainda são um tanto quanto utópicos. Isso não quer dizer que, a cada ferramenta que os profissionais da computação adicionem aos programas, não haja uma evolução nos procedimentos de uso e resultados para todos os usuários.

Alguns centros interdisciplinares, como o NILC (Núcleo Interinstitucional de Lingüística Computacional), na USP de São Carlos, já agregam uma grande quantidade de pesquisadores, de ambas as áreas, trabalhando no desenvolvimento de ferramentas para análise. O NILC, junto com departamentos de Lingüística e Computação de outras universidades (como a UNISINOS e a PUC/SP) promove, anualmente, um encontro dos profissionais ligados às áreas no Workshop em Tecnologia da Informação e da Linguagem Humana[4] (TIL, que pode ser acessado em: www.nilc.icmc.usp.br/til/index.htm). Juntos, muitos profissionais discutem novas possibilidades nas áreas afins.

A área de Lingüística Computacional, como pudemos ver, ainda é incipiente, porém promete ser um grande desafio para os profissionais de ambas as áreas. O desafio se mostra, sobretudo, na necessidade, por parte dos profissionais, de dominar duas áreas distintas do conhecimento e gerir, em grupos multidisciplinares, a enorme quantidade de dados disponíveis.

Bibliografia

BARBOSA, M. A. Dicionário, vocabulário, glossário: concepções. In: ALVES, I. M. (Org.). **A constituição da normalização terminológica no Brasil**. 2 ed. São Paulo: FFLCH/CITRAT, 2001.

[3] Esses programas trabalham com arquivos de glossários que são adicionados à tradução corrente. Isso possibilita que muitos tradutores possam trabalhar a mesma tradução ao mesmo tempo e que uma nova versão de uma tradução seja gerada mais rapidamente, já que as únicas palavras que serão traduzidas são aquelas que ainda não constam da versão anterior do texto.

[4] Esse evento, nas duas edições que já realizadas (2003 e 2004), esteve associado aos congressos da Sociedade Brasileira de Computação. Houve consenso entre os presentes que a próxima edição deveria estar associada a algum congresso de lingüística, mas novamente ela acontecerá em um evento da SBC.

FROMM, G. A construção de sentido em vocabulários técnicos. In: **Revista CROP n. 10**. São Paulo: Humanitas, 2004.
FROMM, G. O uso de corpora na análise lingüística. In: **Revista Factus 1**. Taboão da Serra: FTS, 2003.
FROMM, G. Ferramentas de análise lexical computadorizadas: uma aplicação prática. In: **Revista Factus 3**. Taboão da Serra: FTS, 2004.
NILC (Núcleo Interinstitucional de Lingüística Computacional). Órgão do Instituto de Matemática e Computação da USP de São Carlos. Disponível em <http://www.nilc.icmc.usp.br/nilc/>. Acesso em 12.03.2005.

Guilherme Fromm é professor da Universidade Bandeirante de São Paulo (UNIBAN). Mestre em Lingüística e doutorando em Língua Inglesa, ambos pela USP.