LINGÜÍSTICA COMPUTACIONAL: ELABORAÇÃO DO DIPLOMA PARA A LÍNGUA PORTUGUESA *

José Marcelino Poersch
Departamento de Lingüística – PUCRS

0. SUMÁRIO

Conceitua-se a lingüística computacional como sendo uma ciência interdisciplinar na qual o lingüista se serve de, fornece subsídios a, e interage com a ciência da computação.

Acredita-se que compete ao lingüista conquistar espaços na área da informática para efetuar contribuições ou realizar aplicações. Revistas especializadas apontam para várias destas aplicações, classificadas segundo os diversos níveis da análise lingüística. Assim, a Fonética e Fonologia contribui na elaboração de programas de síntese e de reconhecimento da voz humana, bem como na determinação de regras de separação silábica automática para editores de textos. À Lexicografia compete fornecer listas de freqüencia de palavras para a organização de dicionários bilingües e a programação de "ortografadores" (**spellers**). A Morfossintaxe e a Semântica desempenham papel importante nos programas de tradução automática e de processamento eletrônico de textos. Isto sem falar do papel preponderante do qual os lingüistas foram incumbidos, junto a grupos de investigação das Ciências Cognitivas e da Inteligência Artificial, no desenvolvimento de teorias sobre o processamento da informação lingüística, tanto no aspecto receptivo quanto produtivo.

O presente artigo relata o contexto teórico, metodológico e histórico de uma investigação em lingüística quantitativa visando

* Este artigo é uma versão da comunicação apresentada no I Congresso Brasileiro de Lingüística Aplicada, em Campinas.

a fornecer subsídios aos engenheiros de computação da Apple Computer para a elaboração de um **hardware kit** (conjunto ou dispositivo eletrônico embutido na Unidade Central de Processamento), denominado Diplomata, para microcomputadores Apple IIe, dotando-os, mediante o simples acionar de um comutador, da possibilidade de utilizar, instantaneamente, além do teclado padrão do inglês, de um teclado específico para a edição de textos em língua portuguesa. Esse dispositivo permite gerar todas as letras, com os respectivos sinais de acentuação, mediante o toque de uma única tecla, e exibi-los no visor com uma forma idêntica àquela com que deverão aparecer na versão impressa.

A investigação consistiu no levantamento da distribuição de freqüência de todos os caracteres gráficos utilizados em textos do português e num estudo da distribuição ótima dos mesmos num teclado auxiliar.

A pesquisa foi encomendada ao Centro de Pesquisas Lingüísticas da PUCRS pela International Solutions, firma estabelecida em Sunnyvale, Califórnia, que obteve da Apple Computer a permissão de fabricar o Diplomata, dispositivo que permite gerar e exibir caracteres diferentes daqueles especificados pela International Standard Organization. A característica principal do Diplomata é sua comutabilidade, recurso que permite uma conversão instantânea entre dois ou mais conjuntos de caracteres.

1. INTRODUÇÃO

"É preciso que os homens saibam que (nos tempos modernos) eles não podem nunca parar de aprender. Se você não começar logo em seguida (à obtenção de um diploma, mesmo que seja de uma instituição renomada) a se pôr a par dos novos conhecimentos, ou a aprender a utilizar os novos instrumentos, seu diploma não vale mais nada. É preciso aprender durante toda a vida, porque a revolução científica não vai parar". Se estas palavras de Servan-Schreiber (1985, pág. 4) são verdadeiras para qualquer ramo da ciência, com muito maior razão o são para a área da ciência lingüística a qual, a cada momento, encontra novos horizontes de investigação e de aplicação, constituindo-se numa das ciências mais produtivas da atualidade. E tanto isto é verdade que basta um afastamento de poucos anos para o estudioso sentir-se desatualizado.

Neste avanço contínuo, o saber tem que tornar-se cada vez mais especializado. Um claro exemplo disso é a atual caminhada que a Lingüística enceta, lado a lado, com a Informática. Haja vista o último Instituto Lingüístico desenvolvido pela Sociedade Americana de Lingüística (LSA), em junho e julho de 1986, na City University of New York (CUNY) onde uma grande parte das disciplinas oferecidas exigiam a manipulação do microcomputador, em diversas áreas, como fonologia, morfologia, semântica, sintaxe, psicolingüística, teoria da variação, linguagens da Inteligência Artificial. Considerem-se, entre tantas, as seguintes disciplinas: Semântica Computacional (Sowa), Estatística para análise de dados (Sankoff), Experimentos Psicolingüísticos com uso de computador (Chodorow), LISP para lingüistas (Langsam), Modelos formais e computacionais para a aquisição infantil da gramática (MacWhinney), Abordagens contextuais e computacionais no desenvolvimento da linguagem (Cross), Computadores na tradução (Teller), Seminário em morfologia computacional (Aronoff), Modelos computacionais de aquisição da linguagem (Dogherty), Implementação do computador na análise e síntese da voz (Rubin), PROLOG para lingüistas (Dahlgren).

Com base neste programa, pode-se perceber que a Lingüística Computacional avança com passos de gigante, desbrava novas trilhas, fornece contribuições importantes à informática. É neste contexto que se insere a contribuição relatada no presente artigo.

2. LINGÜÍSTICA COMPUTACIONAL

2.1. Conceituação

Conceituamos Lingüística Computacional (LC), no seu sentido mais amplo, como área interdisciplinar da ciência da linguagem apresentando um tríplice objetivo:

a) servir-se do computador como instrumento de trabalho com o objetivo de processar e analisar seus dados, editar seus textos e controlar seus experimentos;
b) fornecer subsídios à informática para a obtenção de software básico;

c) incentivar uma colaboração mútua com a ciência da computação visando ao progresso dos diversos projetos da Inteligência Artificial.

Da lingüística, a LC explora tanto o aspecto quantitativo quanto o aspecto qualitativo. No aspecto quantitativo, onde pontifica a lingüística estatística, usa-se o computador para o processamento e análise dos dados. "A determinação da extensão daquilo a que o falante está preso pelo código lingüístico e, contrariamente, a extensão daquilo que lhe dá liberdade (pelo que ele é livre) e pelo qual ele pode ser original, é a essência daquilo que se chama de Lingüística quantitativa" (Herdan 1966, p.5 e 6). No aspecto qualitativo, algébrico, de formalização da linguagem, a lingüística fornece, ao cientista da computação, elementos que possibilitam a elaboração de linguagens intermediárias, elementos que permitem a segmentação (parsing) tão necessária à análise automática de textos. A linguagem natural é analisada sob o prisma da Lógica Simbólica (Alwood 1977), em seus ramos da teoria dos conjuntos, do cálculo sentencial, da lógica inferencial (argumentos e inferências) e do cálculo funcional.

2.2. O computador a serviço da lingüística

A LC, num primeiro momento, procura utilizar-se de um computador para processar e analisar estatisticamente os dados quantitativos da língua. Isto realmente constitui a aproximação mais antiga entre a lingüística e a ciência da computação. Com o auxílio do computador, o lingüista consegue:

a) executar automaticamente o levantamento da freqüência de ocorrência dos diferentes elementos e unidades lingüísticas de um texto (e de acordo com diversos níveis): letras, afixos, palavras, sintagmas, categorias gramaticais e, até, frases, com o objetivo de, por exemplo, organizar o vocabulário fundamental ou efetuar a análise estilística de textos de um determinado autor ou de uma determinada época;
b) armazenar e recuperar informação lingüística;
c) extrair dados bibliográficos para um determinado artigo ou determinada obra;

d) implementar pesquisas psicolingüísticas relacionadas principalmente com os processos de recepção, como Tarefas de Decisão e de Acesso Lexical (Lexical access and lexical decision tasks) relatados em diversos números do Journal of Memory and Language de 1985 (Hudson and Bergman 1985, Goodman 1985), com leiturabilidade de textos e com medidas de maturidade lingüística;
e) analisar estatisticamente quaisquer dados de lingüística quantitativa, principalmente da variação lingüística, visto já existirem programas gerais de estatística para a lingüística e programas especiais para a variação lingüística (VARBRULE de Sankoff) adaptados a microcomputadores;
f) usar um editor de textos para produção de monografias, de relatórios, de obras literárias e de material de ensino (incluindo polígrafos e testes);
g) programar exercícios para incrementar a rapidez e melhorar o nível de compreensão de textos.

2.3. A lingüística colabora com a computação

Numa segunda instância, a LC persegue objetivos mais ambiciosos quais sejam os de contribuir diretamente com a ciência da computação na produção de software básico como o design de um teclado ideal para a língua portuguesa baseado em dados de freqüência de ocorrência dos caracteres gráficos e dos encontros ou seqüências destes mesmos caracteres, e na elaboração de um ótimo editor de textos mediante o fornecimento de um algoritmo que possibilite uma separação silábica automática e mediante a montagem de ortografadores que permitem uma correção automática de textos. Necessária se faz a contribuição do fonólogo e do foneticista no sentido de descrever criteriosamente todos os sons e fonemas (em seus formantes) de uma língua para obter a síntese da voz e o reconhecimento da fala natural. A lingüística também tem uma significativa contribuição a emprestar na organização das linguagens de programação, principalmente das mais evoluídas ou das mais próximas da linguagem natural como LISP e PROLOG. Enfim, é à lingüística que compete fornecer aos informáticos informações sobre o processamento (receptivo e produtivo) de dados

da comunicação lingüística para possibilitar o translado de conhecimentos da mente humana ao computador com o objetivo de torná-lo máquina inteligente. Quanto aos aspectos teóricos da lingüística que devem ser do conhecimento de um cientista da computação, Ulf Gregor Baranow (1983, p.23-36) em seu artigo "Perspectivas na contribuição da lingüística e de áreas afins à ciência da informação", faz um estudo bastante minucioso e chega ao ponto de preconizar o estabelecimento de uma disciplina, nos currículos dos cursos de Informática, que trate da natureza da linguagem.

2.4. Interatividade

O terceiro aspecto relevante da LC consiste numa interatividade entre lingüistas e informáticos para resolver problemas mais complexos, problemas relacionados com as duas ciências. Faz-se menção, principalmente, à análise automática de textos e à tradução automática, atividades que pressupõem a colaboração do lingüista para o estabelecimento de algoritmos que consigam distinguir as diversas categorias gramaticais, discriminar as diversas funções das palavras e, principalmente, desambiguar homografias. Para a indexação automática de um texto, o informático, além dos recursos de um bibliotecário, deve também servir-se dos conhecimentos de um lingüista para saber como identificar os dados que merecem ser levantados (Haller 1986).

No campo da Inteligência Artificial, onde pesquisadores exploram as questões básicas da inteligência humana e procuram construir modelos desta inteligência em computadores (Schank & Childers 1984, p.29), é ao lingüista que compete fornecer os dados sobre a linguagem mais condizente para o processamento de dados de sistemas especializados (**expert systems**) que constituem uma das metas prioritárias do importante projeto japonês da 5ª geração (Figenbaum & McCorduck 1984).

> No processamento da linguagem natural, trata-se de desenvolver programas que entendam a linguagem natural falada e escrita, visando à interação com o computador por meio desta mesma linguagem, e não mais em linguagem computacional, o "computês". Para um programa computacional interpretar uma comunicação em linguagem natural, o "conhecimento" necessário envolve a estrutura das sentenças, o significado das palavras, a morfologia das palavras e as regras de conversão, entre outros requisitos. (Visão 1986, p.35.)

2.5. Algumas aplicações

Analisada desta maneira a vasta gama de atividades envolvidas pela LC, ficam ressaltadas as inúmeras aplicações desta disciplina. Citaremos algumas:

a) Colaborar, através de pesquisas quantitativas, no aperfeiçoamento de editores de textos, dotando-os de todos os recursos possíveis no que tange à separação silábica automática, aos ortografadores que permitem uma correção automática, e a dicionários (**tesaurus**) que fornecem dados imediatos ao editor/escritor quanto a sinônimos e significados.

b) Contribuir significativamente na vasta área do Ensino da Linguagem com uso do computador (Costa 1986) mediante a elaboração de programas instrucionais, tanto no âmbito da leitura e da escrita quanto do vocabulário e da gramática. O levantamento do português fundamental será de valiosa contribuição para a elaboração de material didático graduado e adaptado ao nível de maturidade lingüística do aprendiz, quer se trate do português como primeira língua ou como segunda língua.

c) Facilitar a localização de dados lingüísticos os mais diversos, incluindo informação bibliográfica, informação acerca de pesquisas, de cursos, de especialistas, de revistas e de Encontros da área.

e) Possibilitar a produção, na área da tradução, de dicionários bilíngües, de informação simultânea em várias línguas (Projeto EUROTRA) e de programas especiais para munir telefones com interpretação automática instantânea.

f) Propiciar o desenvolvimento de programas de indexação automática de livros.

Considerando todas estas aplicações, verifica-se que o lingüista realmente se encontra frente a uma variedade muito significativa de oportunidades. Acontece que raras vezes o cientista da computação "se dá conta da existência do lingüista. Compete a este descobrir estas oportunidades" (Mazzocco 1979, p.6).

3. O PROGRAMA "DIPLOMATA"

3.1. Contexto histórico

Durante o período em que desenvolvemos nosso programa de Pós-doutorado em Lingüística Cognitiva na Universidade da Califórnia, em Berkeley, em 1985, acompanhando as atividades do Instituto de Ciências Cognitivas, fomos compelido a nos envolver com a Lingüística Computacional e, naturalmente, com os computadores. Entre outros problemas que nos começaram a afligir, o de encontrar um bom editor de textos para a língua portuguesa foi um dos principais. Havíamos adquirido um microcomputador APPLE IIe. Depois de usar, inicialmente, o editor "Janela Mágica", passamos a analisar, sucessivamente, os editores "Apple Works", a versão do Word Star para o CPM e, finalmente, a recente versão do Word Perfect, especialmente adaptado para o APPLE IIe.

Verificamos que todos estes programas apresentavam deficiências quanto aos recursos de acentuação (ou sinais supra e infra grafêmicos). Certos caracteres necessitam do toque de quatro ou cinco teclas para serem produzidos. Outros problemas concorrentes eram: o aparecimento do texto no visor de forma diferente ao da versão final impressa e a separação silábica automática, recurso necessário para obter um texto com justificação na margem direita.

Depois de criteriosas análises, contactamos diretamente com a Apple Computer e reclamamos a falta de um bom editor de textos para a língua portuguesa. Fomos, então, informados de que a Apple Computer havia autorizado uma companhia de Sunnyvale (Califórnia) para a produção de programas especiais para cada língua que apresentassem caracteres gráficos diferentes do inglês. O nome do programa: **The Diplomat.** O nome da firma: **International Solutions.** Diretor de marketing: **Richard Vedder.**

3.2. Design do programa

O "Diplomata" (International Solutions 1983) consiste num dispositivo eletrônico embutido na Unidade Central de Processamento (CPU) de microcomputadores Apple IIe, dotando-os, mediante o simples toque de uma tecla, da possibilidade de utilizar,

instantaneamente, além do teclado padrão para o inglês (Standard ANSI Keyboard), de teclados específicos para a edição de textos em línguas diferentes. A característica básica do "Diplomata" é sua comutabilidade, qualidade que permite uma conversão instantânea para dois ou mais conjuntos de caracteres. Este dispositivo permite gerar todas as letras e demais caracteres gráficos (como sinais de acentuação) mediante o simples toque de uma única tecla e exibi-los no visor com uma forma idêntica àquela com que deverão aparecer na versão impressa.

De posse desses dados fomos visitar a firma produtora a fim de obter informações complementares. O diretor gentilmente nos mostrou o programa e o rodou no microcomputador. Tivemos, então, uma idéia das possibilidades do programa, de sua constituição e de seu acessamento. Verificamos que a distribuição dos caracteres no teclado apresentava as seguintes falhas ou deficiências: permutação desnecessária e injustificável de letras (A pelo Q, W pelo Z), inclusão de caracteres não pertencentes ao alfabeto do português (W, Y, K), exclusão de algumas vogais acentuadas (í e á), privilegiamento de certas vogais acentuadas (à) em detrimento de outras (á, ê e ó) bem mais produtivas. Diante desta constatação, inquirimos o diretor a respeito do critério adotado para semelhante distribuição. A resposta foi bastante lacônica: "Foi com base em informações recebidas de pessoas falantes do português". Deduzimos desta resposta que não havia sido utilizada informação suficientemente científica baseada num levantamento da freqüência de uso desses caracteres em língua portuguesa. Apresentamo-nos, então, como lingüista e oferecemos nossos préstimos para a realização de uma pesquisa que visasse ao levantamento de dados quantitativos que servissem de base para a distribuição dos caracteres gráficos num teclado auxiliar. Aceita a proposta, assinamos um protocolo de intenções.

3.3. Objetivo da pesquisa

De volta a PUCRS, em março de 1986, incluímos esta pesquisa no programa geral de metas e pesquisas do **Centro Brasileiro de Lingüística Computacional.** Este centro desenvolve um programa interdisciplinar de pesquisas contando com a colaboração de especialistas em computação, em estatística e em educação (área de en-

sino e aprendizagem da linguagem), sob a nossa direção. O centro atua junto ao curso de doutorado em lingüística. Inserido no programa geral, existe um projeto setorial – Editor de Textos – que objetiva prover a informática de dados lingüísticos confiáveis para a elaboração de bons editores de textos que se constituem em programas dotados de uma ótima disposição dos caracteres gráficos no teclado, de um perfeito separador silábico, peça indispensável para executar a justificação da margem direita de um texto, de um bom ortografador que permite uma correção automática de, pelo menos, 70% dos erros ortográficos, e de um **tesaurus** que proveja pronta e ampla informação relativo a significados, sinônimos e categorias gramaticais.

Fundamentamos a pesquisa, quanto ao referencial teórico e quanto aos parâmetros metodológicos, em Zipf (1949), em Guiraud (1959) e em Herdan (1966).

Em Guiraud (1959, p.31) lemos que

> a linguagem é um sistema de signos e, como tal, é submetida às leis das probabilidades; ... A freqüência dos diferentes fonemas é estabelecido sobre um compromisso entre a economia da transmissão e aquela da recepção; assim a redação de um telegrama tende para o menor número de palavras compatível com a compreensão da mensagem. A freqüência não tem, portanto, de forma alguma, um caráter arbitrário; ela é determinada pela função, pela natureza do signo e pelas suas coordenadas físico-psicológicas.

Sendo que, no sistema escrito, as letras mantêm correspondência com os sons, fácil fica concluir que a freqüência daquelas está sujeita às leis probabilísticas.

Herdan (1966), p.15) afirma que

> as proporções das formas lingüísticas pertencentes a um nível particular de compreensão, ou a um estágio de decodificação lingüística, – fonológica, gramatical, semântica – permanece sensivelmente constante para uma dada língua, num dado tempo de seu desenvolvimento e para um número suficientemente grande de observação.

Como as letras não estão diretamente ligadas a um significado e, portanto, não dependem da variável escolha individual, os dados estatísticos de Zipf comprovam a constância de sua distribuição em amostras das mais variadas.

Para a presente pesquisa tomaram-se como amostra duzentas mil palavras de artigos científicos, já que estes constituem o produto mais representativo da utilização de editores de textos. Estes textos foram digitados num microcomputador e, posteriormente, processados por um programa especialmente escrito para tal propósito. O resultado foi uma lista da freqüência dos caracteres gráficos da língua portuguesa, ordenados segundo sua freqüência percentual.

Partindo da realidade do teclado do Apple IIe, que apresenta 48 teclas possibilitando 96 caracteres em sua caixa alta e baixa, fizemos uma nova distribuição dos caracteres em textos de artigos científicos abrindo, desta maneira, espaço para os caracteres da língua portuguesa ausentes no teclado standard do inglês. Os novos caracteres foram localizados segundo critérios de comodidade e de produtividade.

4. CONCLUSÃO

Acreditamos que, desta maneira, oferecemos uma valiosa e significativa contribuição aos técnicos da informática, baseados em dados empíricos da lingüística. Usando informação lingüística para resolver problemas práticos da linguagem, fica a atividade aqui relatada inserida no basto campo da lingüística aplicada.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ALLWOOD, Jeans et alii. **Logic in linguistics.** Cambridge, Cambridge University Press, 1977.
2. BARANOW, ulf Gregor. Perspectivas na contribuição da lingüística e de áreas afins à Ciência da Informação. **Ciência da Informação.** Brasília, CNPq/IBICT, 12(1): 23-35, 1983.
3. COSTA, Miriam Solange. O computador no ensino de línguas: retrospecto e perspectivas. **Interação.** São Paulo, Difusão Nacional do Livro, 3(18): 17-20, abr. 1986.
4. FEIGENBAUM, Edward and MCCORDUCK, Pamela. **The fifth generation: artificial intelligence and Japan's computer challenge to the world.** New York, New American Library, 1984.
5. GORDON, Barry. Subjective frequency and lexical decision latency function: implications for mechanisms of lexical access. **Journal of Memory and Language, 24**(6): 631-45, dec. 1985.

6. GUIRAUD, Pierre. **Problèmes et methodes de la statistique linguestique.** Dordreche, D. Reibel Publishing Company, 1959.
7. HALLER, Johann. Análise lingüística e indexação automática de textos. **Veritas.** Porto Alegre, PUCRS, 31(123): 393-414, 1986.
8. HERDAN, Gustav. **The advanced theory of language as choice and chance.** Heildelberg, Springer-Verlag, 1966.
9. HUDSON, Patrick and BERGMAN, Jarijke. Lexical knowledge in word recognition: word length and word frequency in naming and lexical decision tasks. **Journal of Memory and Language,** 24: (1): 46-58, feb. 1985.
10. INTERNATIONAL SOLUTION. **The Diplomat: installation manual.** Firth edition. Saratoga (Ca), International Solutions, 1983.
11. MAZZOCCO, Alexis (entrevista). Opportunities for linguists in the field of computers. **The linguistic reporter.** sep. 1979.
12. SANKOFF, David and CEDERGREN, Henrietta (eds.). **Variation omnibus.** Alberta, Linguistic Research, 1981.
13. SCHANK, Roger & CHILDERS, Peter. **The cognitive computer: on language, learning and artificial intelligence.** Menlo Park, Addison-Wesley Publishing Company, 1984.
14. SERVAN-SCHREIBER, Jean-Jacques (entrevista). Informática e informação. **Veja.** São Paulo, Editora Abril, (900): 3-5, 4 dez. 1985.
15. Sociedade Americana de Lingüística. **1986 Linguistic Institute.** New York, CUNY, 1986.
16. Visão (autor não citado). Inteligência artificial: o Brasil entra na corrida. **Visão.** p.34-38, 22 jan. 1986.
17. ZIPF, G. K. **Human behavior and the principle of least effort.** Cambridge (Mass.), Addison-Wesley Publishing Company, 1949.