REDISTRIBUIÇÃO DOS CARACTERES GRÁFICOS NO TECLADO DE MICROCOMPUTADORES COM BASE NA LINGUÍSTICA QUANTITATIVA

José Marcelino Poersch*

Através do estudo da frequência dos caracteres gráficos em Língua portuguesa em combinação com o estudo dos reflexos dos dedos na digitação e do esforço para acessar as diversas teclas, são sugeridas mudanças no actual teclado QWERTY de microcomputadores e de outras máquinas eletrônicas de escrever com o objectivo de veicular, na digitação, o máximo de informação com o mínimo de custo.

O presente estudo reavalia a afirmação de linguístas de que a tipologia do texto não constitui variável na frequência dos caracteres, visto não estar a distribuição dessa frequência vinculada com o significado. Por outro lado, oferece ao analista dados do português que podem ser comparados com dados de outros idiomas, notadamente do inglês, do francês e do alemão.

Essa nova distribuição, por certo, enfrentará o conservadorismo exagerado que obstaculiza a promoção dos avanços científicos e tecnológicos no mundo cultural de maneira semelhante à lei da inércia que dificulta

* PUCRS.

mudanças de movimento no mundo físico. No entanto, se no dia de amanhã uma pesquisa experimental vier provar que essa nova distribuição permite digitadores mais velozes do que o oportunizado pelo teclado QWERTY, é de se supor que a tecnologia, num futuro não muito distante, passe a produzir esse novo teclado.

1. Os teclados de microcomputadores

Com o avanço da tecnologia, procuram-se instrumentos cada vez mais perfeitos, mais ágeis e mais fáceis de serem operados. Constata-se que os teclados dos microcomputadores, em sua versão padrão, apresentam dificuldades para a produção de textos em português, principalmente no que se refere a caracteres gráficos específicos quais sejam a cedilha e as vogais acentuadas (em número de 12). Diversas tentativas de solução foram propostas, experimentadas e integradas aos editores de texto, programas responsáveis pelos recursos que um microcomputador pode oferecer para editar um texto. Na maioria dos casos, utiliza-se a sobreposição de caracteres (o c cedilha é obtido sobrepondo uma vírgula ao c (ç = c + ,) mediante o toque de três, quatro ou cinco teclas para a produção de um desses caracteres específicos (ã: letra a + utilidades + sobreposição + caixa alta + caracter (~)). Além desse exagerado consumo de energia (tempo), ainda existe o inconveniente de o visor nem sempre representar o caracter da maneira como ele deve aparecer na impressão, ou o visor registar mais caracteres do que deverão aparecer na impressão o que prejudica enormemente o controle do emparelhamento da margem direita do texto. Outros aparelhos, por exemplo, por possuírem teclas para funções especiais, podem ser programados para produzir esses caracteres, mediante um retrocesso, como nos teclados tradicionais das máquinas datilográficas. No entanto, sempre será mais de dois toques para produzir uma letra acentuada. Isso gera um consumo desnecessário de esforço.

Uma firma californiana, sediada em Sunnyvale, recebeu da Apple Computer a autorização para produzir o "Diplomata", programa especial

que dota os editores de texto da possibilidade de produzir caracteres específicos de línguas diferentes do inglês. O "Diplomata" consiste num dispositivo eletrônico embutido na unidade central de processamento (CPU) dos 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), teclados específicos para a edição de textos em línguas diferentes. A característica básica do "Diplomata" é a sua comutabilidade, qualidade que permite uma conversão instantânea entre dois ou mais conjuntos de caracteres. Esse dispositivo permite gerar todas as letras bem como os 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.

Analisando esse programa, verifica-se que a distribuição dos caracteres no teclado apresentam várias falhas ou deficiências: permutação desnecessária e injustificável de letras (A pelo Q, W pelo Z), da inclusão de caracteres não pertencentes ao alfabeto do português (K, W, Y), da exclusão de algumas vogais acentuadas (í, á), do privilegiamento de certas vogais acentuadas (à) em detrimento de outras (á, ê, ó) que são bem mais produtivas.

As deficiências do "Diplomata" trouxeram a ideia de proceder a um levantamento criterioso da frequência dos caracteres gráficos em português para, mediante a eliminação de caracteres desnecessários ou de baixa produtividade na editoração de textos em língua portuguesa, abrir espaço para a inserção de outros caracteres específicos da nossa língua e dispô-los segundo sua frequência de ocorreência.

A par deste problema, verifica-se um outro: a disposição dos caracteres gráficos num teclado padrão.

Pergunta-se quais foram os critérios que nortearam os engenheiros de máquinas de escrever a distribuir os caracteres gráficos pelo teclado, na forma como eles se encontram dispostos atualmente. Sabe-se que essa dis-

posição deve obedecer a certos critérios que devem levar em conta a distância em que as teclas se localizam dos dedos da mão posicionada (em seu ponto de repouso) na parte central da segunda carreira de teclas, de baixo para cima, e a energia diferenciada que cada dedo possui (o indicador, o médio, sabidamente, possuem mais energia do que o anular e o mínimo). Partindo do pressuposto de que os caracteres mais produtivos devem corresponder à posições mais fáceis de serem acessadas, fácil se torna deduzir que o levantamento de frequência dos caracteres gráficos muito pode contribuir para a elaboração dessa distribuição.

O teclado de um microcomputador é igual ao de uma máquina de escrever eletrônica padrão (teclado QWERTY), acrescido de alguns outros comamdos. O teclado é assim denominado devido à disposição das seis primeiras letras do teclado superior. Essa disposição dos caracteres do teclado não foi a primeira a ser elaborada e utilizada e, como todas as outras disposições, levava em conta a frequência de ocorrência dos caracteres da língua inglesa. Os outros teclados, contudo, não se perpetuaram porque as letras mais frequentes e de ocorrência sucessiva (tais como D e E) estavam dispostas lado a lado no teclado e, devido à rapidez de toque dos datilógrafos, as teclas se acavalavam na hora de atingir a folha. Assim, para diminuir a rapidez de toque, bem como para evitar danos às máquinas de escrever, Glidden, Sholes e Soulé, três norte-americanos, separando justamente os caracteres de maior frequência de ocorrência, criaram o conhecido teclado QWERTY.

Colocam-se, então, duas questões que devem merecer toda a nossa consideração. Como pode um editor de textos para português bem servir a um digitador se possuir um teclado baseado na frequência dos caracteres da língua inglesa? E como pode-se aceitar o fato de tais caracteres estarem dispostos de forma a aumentar o espaço de tempo entre um toque e o toque imediatamente seguinte a fim de evitar o acavalamento das teclas, se um microcomputador funciona por impulsos elétricos, independente da sequência ou proximidade de alavancas mecânicas?

Com base na problemática assim estabelecida supõe-se que o levantamento da frequência dos grafemas do português, bem como a frequência dos digramas (sequência de dois grafemas) pode conduzir-nos à elaboração de uma disposição mais racional, ou melhor, mais eficiente do teclado se se fizer corresponder às teclas mais rápidas de serem acionadas os caracteres gráficos mais produtivos da língua, de modo a permitir um ganho razoável de tempo.

A análise destes dois aspectos leva-nos a concluir que a linguística pode trazer uma significativa contribuição à tecnologia da ciência da computação através de uma pesquisa que objetive estudar a frequência dos caracteres gráficos e dos digramas do português. E o que acentua ainda mais a interdisciplinariedade é o facto de que o linguista se utiliza do cientista da computação para a elaboração do programa de levantamento de dados e de um microcomputador para o efetivo levantamento automático dos dados a partir de um corpus.

2. Aspectos quantitativos da linguagem

As observações que fazemos na vida comum, bem como as observações mais sistemáticas da ciência, revelam certas regularidades na natureza.

As leis da ciência são formulações que expressam essas regularidades com a máxima precisão possível. Entretanto, nem todas as leis da ciência são universais, pois, em vez de afirmar que tal regularidade ocorre em todos os casos, algumas leis afirmam que só ocorre em certa percentagem de casos. Se a percentagem for especificada ou se se formular uma afirmação quantitativa sobre a relação entre um evento e outro, então formula-se uma lei estatística.

Às vezes, quando damos uma explicação dos fatos, as únicas leis que se aplicam são leis estatísticas e não leis universais. Na falta de leis universais conhecidas, muitas vezes as explicações estatísticas são o único tipo de explicação disponível.

Se, em determinado campo, encontramos ordem suficiente para fazer comparações e afirmar que, em algum aspecto, uma coisa está acima de outra, então, em princípio, há possibilidade de medição. O primeiro passo consiste em formular, regras de comparação e, depois, se for possível, regras quantitativas.

A quantificação dos conceitos leva, antes de tudo, à formulação de um vocabulário mais eficiente; com isso elimina-se a necessidade de recorrer a um vocabulário muito vasto e sujeito a todos os matizes do subjetivismo, com suas variações às vezes imprevisíveis.

O mais importante, porém, é que os conceitos quantitativos nos permitem formular leis quantitativo-explicativas dos fenômenos e permitem a previsão de novos fenômenos ou estados de fenômenos. O enfoque estatístico é especialmente útil quando devemos analisar grande quantidade de dados.

O que a pessoa diz, segundo Miller (1951), encontra-se restringido ou controlado de diversas maneiras: pelo público, pela gramática da língua, por suas próprias necessidades e experiências. Antes de começar a considerar algumas dessas restrições particulares impostas ao que a pessoa diz, necessitamos, sem dúvida, contar com uma orientação estatística geral que nos indique que classe de expressões deve ser levada em conta.

Um tipo de restrição imposta ao falante é a estrutura da língua que utiliza. A fala e a escrita (conduta verbal) estão abertas a todas as influências que afetam qualquer tipo de conduta, e uma análise do contexto completo de um acto comunicacional deve incluir as necessidades do falante, suas percepções, o público com que conta e sua bagagem cultural.

Para controlar a amplitude do contexto verbal e avaliar quantitativamente sua influência, é necessário possuir dados estatísticos referentes às frequências relativas de ocorrência das unidades verbais.

Segundo Malmberg (1971), todo tratamento científico dos fenómenos linguísticos, toda conclusão a respeito deles, toda descrição de estados de língua pressupõem, talvez em sua forma mais simples, o auxílio da estatística.

A imagem da língua só é completa se pudermos medir a frequência dos tipos e das combinações; por um lado, no vocabulário tal como ela aparece no dicionário e, por outro lado, na língua viva, ou nos textos impressos. Assim, a descrição de um determinado fonema de um sistema fonológico e a descrição de suas possibilidades de combinação com outros fonemas deve ser completado por uma pesquisa sobre a frequência dos ditos elementos e das combinações em questão, comparada à de outros fonemas. A descrição qualitativa deve ser completada por uma pesquisa quantitativa.

O desenvolvimento humano e o avanço das civilizações dependeram, além de outras variáveis, do progresso alcançado nos meios de receber, de comunicar e de registar o conhecimento.

A comunicação, segundo Cherry (1971), implica essencialmente uma linguagem, um simbolismo, quer seja um dialeto falado, uma inscrição em pedra, um sinal do Código Morse, ou uma série de impulsos binários num computador moderno.

As formas de comunicação sofreram inúmeros processos de mudanças desde os pictogramas, ideogramas e hieroglifos, até a sua forma atual. Neste processo de mudança, sempre imperou a preocupação com a economia dos símbolos usados, procurando evitar a redundância, sem, contudo, descurar a eficiência da comunicação

Com a introdução do código MORSE, percebeu-se o aspecto estatístico da economia da linguagem. Morse se deu conta de que as várias letras da língua inglesa não são usadas com igual frequência; uma visita a uma tipografia e uma contagem das quantidades de tipos usados forneceu-lhe uma estimativa das frequências relativas das letras. Morse concebeu o

seu código de modo que as letras mais usadas correspondessem aos símbolos de ponto-e-traço mais curtos.

A lei da frequência de letras de Morse exibe uma tendência, mas é puramente descritiva; a forma precisa dessa e de outras relações foi explorada sistematicamente por Zipf (1949), que fez uma coleta e estudo de aspectos estatísticos da fala e da escrita. Ele foi um dos precursores do tratamento estatístico dos fenômenos estatísticos. Ele mostrou que a complexidade dos fonemas é inversamente proporcional à sua frequência e que os fonemas surdos, nas línguas onde existe a distinção surdo/sonoro, são duas vezes mais frequêntes que os sonoros. Também foi Zipf quem estabeleceu o princípio do "menor esforço" que ele considerou válido não só para os sons mas também para outros elementos da língua, particularmente para as palavras.

Tais resultados tendem a mostrar que a frequência dos fonemas segue leis determinadas, que são função dos seus caracteres físicos (acústicos e fisiológicos) e da reação do ouvido a esses estímulos. Esses estudos tem demostrado que os fonemas não se combinam em unidades superiores por obra do acaso, mas em virtude de princípios determinados.

Segundo Malmberg (1971), tão importante quanto o conhecimento dos diversos fonemas que aparecem no vocabulário ou na língua viva é o das leis que, em cada língua, condiciona a construção de unidades superiores: sílabas, morfemas, palavras.

Inúmeras análises estatísticas de línguas faladas e, em escala bem menor, de línguas escritas, são de interesse dos linguistas, psicólogos e engenheiros da comunicação. Estas análises do comportamento linguístico oportunizam o surgimento de leis definidas.

Atualmente, a análise estatística é um importante método de estudo linguístico. Além da frequência de letras e palavras, já foram feitas

observações acerca da frequência de sílabas, de partes da oração, dos hábitos de acentuação e inflexão de interlocutores ao telefone.

O conhecimento da frequência dos fonemas é de importância capital para fins práticos. Os primeiros a chamar a atenção dos linguistas para a frequência relativa dos fonemas, ou das letras, foram os estenógrafos.

Em Guiraud (1959, pág. 31) lemos que "a linguagem é um sistema de signos e, como tal, é submetido às leis das probabilidades; ... A frequência dos diferentes fonemas é estabelecido sobre um compromisso entre e economia da transmissão e aquela da recepção; assim a redação de um telegrama tende para o menor número compatível com a comprensão da mensagem. A frequência 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, num sistema escrito, as letras mantêm correspondência com os sons, fácil fica concluir que a frequência daquelas também se sujeita às leis probabilísticas. (Pocrsh, 1986).

Herdan (1966, pág. 15) afirma que "as proporções das formas linguísticas pertencentes a um nível particular de compreensão, ou a um estágio de codificação linguística (fonológica, gramatical, métrica) permanecem sensivelmente constantes para uma dada língua, num dado tempo de seu desenvolvimento e para um número suficientemente grande de observações".

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.

O fato marcante da estabilidade das frequências relativas dos símbolos parece ser uma característica comum das formas linguísticas. Segundo Herdan (1966, pág. 16) "existe uma ampla similitude entre os

membros da comunidade falante, não apenas quanto ao sistema fonémico, ao vocabulário e à gramática, mas também quanto à frequência de uso de fonemas particulares, itens lexicais e determinadas formas e estruturas gramaticais; em outras palavras, uma semelhança não só no que é usado mas também no quantas vezes é usado".

3. Especificação do problema

3.1. Estabelecimento dos objectivos

O objectivo operacional básico (imediato) é levantar, em textos escritos do português do Brasil, a frequência dos caracteres gráficos e dos digramas em posição inicial, medial e final de palavras.

Os objectivos aplicativos (mediatos) são os seguintes: sugerir mudanças no atual teclado padrão QWERTY de microcomputadores e de outras máquinas eletrônicas de escrever tomando em consideração a redução de esforço e o aumento de rapidez; fornecer dados confiáveis aos linguístas aplicados para a elaboração científica de cartilhas de alfabetização; reanalisar e reavaliar a afirmação de linguistas que pressupõem que o tipo de texto não constitui variável (interveniente) na frequência de grafemas e de digramas, visto estes não estarem relacionados com o significado; comparar os dados levantados no português com os dados encontrados em outros idiomas, notadamente no inglês, no francês e no alemão.

3.2. Formulação das hipóteses de trabalho

3.2.1. Os carateres gráficos em textos de língua portuguesa apresentam percentagens de frequência diferentes. Esta hipótese será avaliada com base na listagem da frequência percentual dos caracteres.

3.2.2. O tipo e o tamanho das amostras não influi na distribuição da frequência dos caracteres gráficos. Esta hipótese será analisada com base no coeficiente de correlação calculado entre as ordens percentuais dos caracteres em textos diferentes;

3.2.3. Certos digramas são mais produtivos do que outros. Esta variação será estudada em três posições (inicial, medial e final) e será apreciada com base no percentual das frequências obtidas.

3.3. Operacionalização das variáveis

A variável independente corresponde aos caracteres gráficos de textos em língua portuguesa, classificados como segue: caracteres grafêmicos, caracteres supragrafêmicos, caracteres intergrafêmicos, caracteres numéricos e outros caracteres.

Os caracteres grafêmicos formam dois grupos: consonantais e vocálicos. Os consonantais são: b, c, ç, d, f, g, h, j, k, l, m, n, p, q, r, s, t, w, v, x, y, z. Os vocálicos são: a, e, i, o, u. Os caracteres supregrafêmicos são: acento grave (`), acento agudo ('), acento circunflexo (^), Til (~) e trema ("). Os caracteres supragrafados são: á, à, â, ã, é, ê, í, ó, ô, õ, ú, ü. Entre os caracteres intergrafêmicos, relacionam-se os seguintes: vírgula (,), ponto (.), ponto-e-vírgula (;), dois pontos (:), ponto de interjeição (!), ponto de exclamação (?), barra (/), abrir parênteses (), fechar parênteses ()), ifem (-), aspas ("), apóstrofe ('), travessão (_) e ponto de reticência (...).

A variável dependente corresponde à frequência de ocorrência dos diversos tipos de caracteres. A frequência relativa servirá para ordená-los, isto é, fornecer-lhes um número de ordem (rango). A ordenação será analisada em diversos agrupamentos os quais serão selecionados de acordo com os objectivos propostos.

A variável interveniente corresponde ao tipo de texto, ao assunto tratado e ao tamanho da amostra. Para calcular a influência dessa variável interveniente, calcular-se-á o coeficiente de correlação dos dados de diversas amostras entre si, tomados dois a dois: tipos diferentes de discurso, assuntos diferenciados, texto curto (porém completo) X parte igual de um texto maior, amostra inicial, medial e final de textos, conjunto de textos curtos: provérbios e dizeres de pára-choque de caminhão.

A análise da variável interveniente tem por objectivo avaliar a hipótese (3.2.2.) de que o tipo da amostra não influi no resultado porque os caracteres gráficos independem de significado. Também existe a possibilidade de comparar esses dados com os dados obtidos por outros pesquisadores, principalmente, por aqueles que trabalharem com sistemas linguísticos diferentes do português.

4. Implementação da pesquisa

4.1. População e amostra

O universo da pesquisa inclui todos os textos escritos em língua portuguesa no Brasil. Baseado em pesquisas correlatas (Malmberg, 1971; Guiraud, 1959; Herdan, 1966; Miller, 1951), sabe-se que uma amostra suficientemente significativa está por volta de duzentas mil palavras de texto corrido; isso corresponde, aproximadamente a trezentas e trinta páginas datilografadas em espaço simples, com sessenta colunas. Em termos de caracteres, corresponde a um milhão e duzentos mil.

Embora pesquisadores ligados a esse campo, baseados no pressuposto de que os grafemas e suas sequências diádicas não estão relacionados com o significado, afirmem que não existem variáveis intervenientes a influir no resultado final das frequências, pretende-se reavaliar essa afirmativa. Para tanto, a amostra global foi constituída de amostras parciais que cobri-

ram os aspectos de: tipológia discursiva, estilo utilizado, assunto tratado, objectivo proposto e autor, entre outros.

A amostra global foi constituída de doze amostras parciais, totalizando 437.719 caracteres; portanto, um terço da amostra acima. Se, no entanto, os coeficientes de correlação apresentados pelas frequências das diversas amostras entre si, bem como pelas frequências de cada uma com as frequências totais, forem significativamente elevados, teremos comprovado a suficiência da amostra. No caso contrário, ter-se-á a necessidade de ampliar o corpus inicial.

As doze amostras parciais foram as seguintes:

I Discurso dissertativo correspondendo a um artigo científico completo;
II Discurso literário em verso;
III Discurso literário, em prosa, parte final de um texto;
IV Discurso literário, em prosa, parte inicial de um texto;
V Discurso literário, em prosa, parte medial de um texto;
VI Discurso dissertativo, parte de um texto técnico-científico;
VII Discurso literário, em prosa, dirigido a leitores infantis;
VIII Discurso descritivo-expositivo, material instrucional (didático), sobre geografia;
IX Discurso expositivo, amostras feitas com minitextos, abordando os mais variados assuntos;
X Discurso narrativo-jornalístico, informativo;
XI Texto técnico-científico (expositivo), sobre filosofia;
XII Discurso expositivo de idéias, produção infantil.

4.2. Preparação do corpus e levantamento dos dados

O corpus total foi constituído de doze corpora para atender as explicitações da segunda hipótese (3.2.2). Esse corpus foi digitado num TK 3000 e os dados foram levantados automaticamente, primeiro para cada

amostra parcial, depois para o corpus total. Para tanto foi utilizado um programa especialmente elaborado em linguagem C.

4.3. Apresentação dos resultados primários

Pela aplicação dos programas de contagem dos caracteres gráficos e dos diagramas sobre os doze textos digitados (amostras) verificou-se que o número total de caracteres analisados foi de 437.719, distribuídos por diversos tipos, conforme se pode verificar na Tabela 1.

Tabela 1 - Computação geral dos caracteres digitados nas doze amostras utilizadas

Caracteres Amostras	Total de caracteres	Grafemas	Inter-Grafemas	Algarismos	Outros Caracteres	Espaços
I	76 143	55 529	2 183	731	2 965	16 327
II	40 051	29 993	937	0	972	8 495
III	39 108	28 757	1 510	10	1 630	8 135
IV	38 104	28 458	1 168	30	1 232	7 849
V	38 383	28 652	1 271	8	1 401	7 760
VI	20 212	14 889	741	451	1 327	3 698
VII	39 220	28 951	1 556	11	1 602	8 081
VIII	39 301	29 626	1 103	261	1 427	7 652
IX	38 792	27 913	1 477	16	1 599	8 703
X	38 642	29 765	1 160	263	1 443	7 832
XI	38 569	29 016	1 326	143	1 525	7 454
XII	11 743	8 699	308	27	424	2 445

A seguir relacionaremos os digramas mais frequentes, consideradas suas posições iniciais, mediais e finais. Esses digramas aparecem elencados em ordem decrescente de suas frequências.

Tabela 2 – Os trinta digramas iniciais mais frequentes

Digra-mas	F	Digra-mas	F	Digra-mas	F
DE	4088	MA	1245	NÃ	801
CO	2937	PO	1209	IN	778
QU	2863	NO	1194	EM	765
SE	2473	RE	1142	SO	705
DO	1592	CA	1120	OS	654
ES	1481	PE	1033	EN	636
PA	1404	TE	989	AS	616
UM	1369	DI	919	VE	613
DA	1309	ME	900	FA	556
PR	1258	NA	875	SU	536

Tabela 3 – Os trinta digramas mediais mais frequentes

Digra-mas	F	Digra-mas	F	Digra-mas	F
DE	3652	PO	926	DI	675
CO	2175	MA	905	OS	650
QU	2140	NO	898	IN	580
SE	2035	RE	888	AS	572
DO	1447	CA	766	EN	507
UM	1225	NA	765	FA	446
DA	1172	EM	743	SO	431
ES	1093	PE	730	SU	416
PA	1033	ME	701	VE	407
PR	999	TE	700	OU	401

Tabela 4 – Os trinta digramas finais mais frequentes

Digra-mas	F	Digra-mas	F	Digra-mas	F
DE	3783	TE	1459	OU	925
OS	3721	IA	1439	TA	888
DO	3306	SE	1433	UM	871
AS	3111	ES	1321	NA	837
AO	2641	AR	1205	RO	760
EM	2074	ER	1170	AL	741
UE	2058	MA	1089	CA	723
DA	1941	IS	988	MO	701
RA	1878	NO	936	AM	684
TO	1696	OR	929	EU	650

4.4. Tratamento estatístico

Os caracteres gráficos foram subdivididos em cinco categorias: caracteres grafêmicos, caracteres intergrafêmicos, caracteres numéricos, outros caracteres e espaços. Os caracteres supragrafêmicos não são computados nessa relação por se apresentarem acavalados com os caracteres grafêmicos (vocálicos). As frequências absolutas e percentuais desses caracteres gráficos se encontram na Tabela 5.

Tabela 5 – Distribuição geral das frequências dos caracteres gráficos em língua Portuguesa

	Frequências	
Caracteres	**Absolutas**	**Percentuais**
Grafêmicos: Total	324.951	72,69
Consoantes	167.965	51,69
Vogais	156.986	48,31
Intergrafêmicos	14.317	3,20
Numéricos	1.719	0,38
Outros	16.151	3,62
Espaços	89.890	20,18
Total	447.028	

Verifica-se que, entre os caracteres grafêmicos, as consoantes, embora tipicamente sejam mais numerosos do que as vogais (22 para 5), sua ocorrência total somente apresenta uma frequência de 51,69% contra 48,31% das vogais.

As tabelas a seguir (6, 7 e 8) fornecem as frequências absolutas e percentuais, respectivamente, dos caracteres grafêmicos, dos caracteres supragrafêmicos e dos caracteres intergrafêmicos.

Tabela 6 – Frequência dos caracteres grafêmicos no corpus total

Identificação	F	%	Postos
A	44.400	13,66	1
B	3.290	1,01	19
C	11.968	3,68	12
Ç	1.819	0,55	21
D	16.700	5,14	8
E	40.446	12,45	2
F	3.500	1,08	18
G	4.321	1,33	16
H	3.208	0,99	20
I	21.808	6,71	5
J	732	0,22	24
K	48	0,01	26
L	9.209	2,83	13
M	14.612	4,50	10
N	17.481	5,38	7
O	35.867	11,04	3
P	8.498	2,61	14
Q	3.545	1,09	17
R	21.406	6,59	6
S	25.307	7,79	4
T	14.987	4,61	9
U	14.465	4,45	11
V	4.886	1,50	15
W	137	0,04	25
X	878	0,27	23
Y	32	0,01	27
Z	1.401	0,43	22

Tabela 7 – Frequência dos caracteres supragrafémicos no corpus total

Identificação	F	%	Postos
Á	1.334	13,96	3
À	228	2,38	9
Â	170	1,77	11
Ã	2.762	28,91	1
É	1.749	18,31	2
Ê	637	6,66	6
Í	1.081	11,31	4
Ó	698	7,30	5
Ô	88	0,92	12
Õ	273	2,85	8
Ú	315	3,29	7
Ü	217	2,27	10

As tabelas 6, 7 e 8 bastam, por si só, para confirmar a primeira hipótese (4.2.1): os caracteres gráficos, em textos de língua portuguesa apresentam diferentes percentagens de frequência. Devido a essa diversidade de frequências, os caracteres podem ser relacionados em ordem decrescente. Entre os caracteres grefêmicos destacam-se os vocálicos pela sua alta frequência: A (13,66%), E (12,45%), O (11,04%) I (6,71%) e U (4,45%).

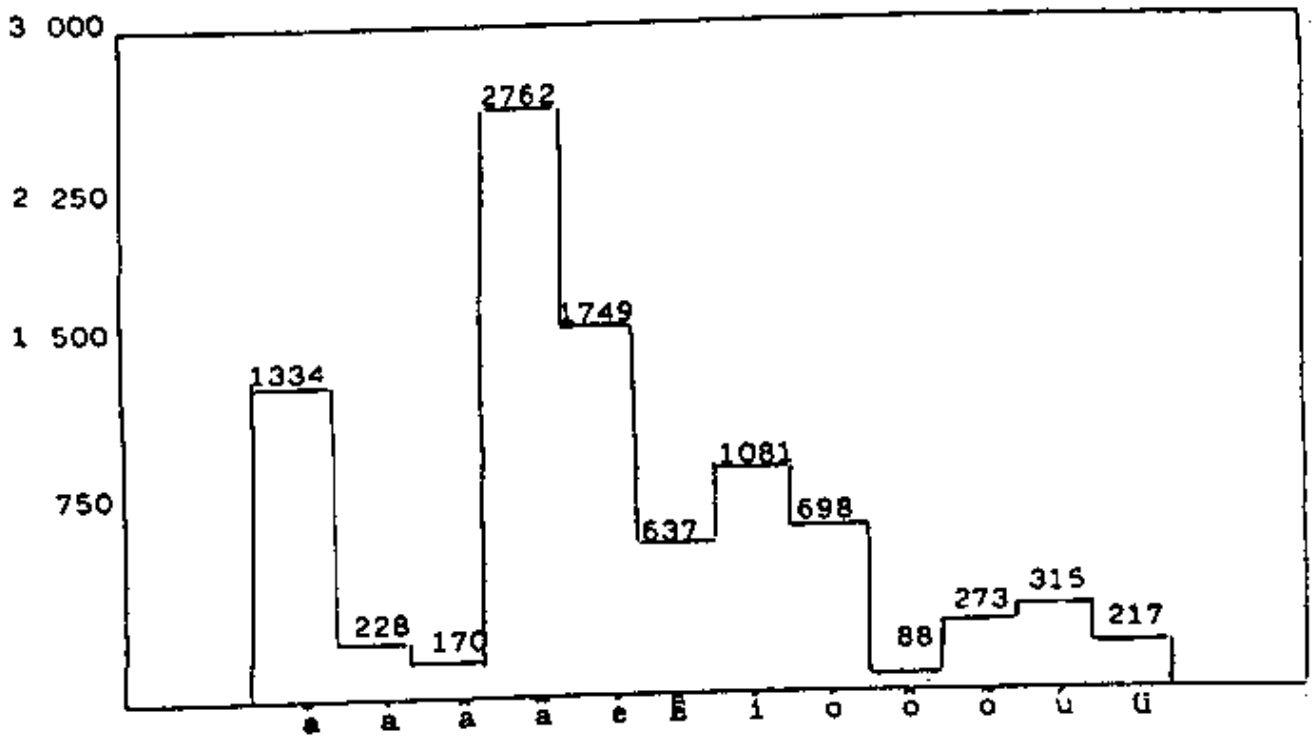
Tabela 8 – **Frequência dos caracteres intergrafêmicos no corpus total**

Identificação	F	%	Postos
,	5.015	35,02	1
.	4.223	29,49	2
;	243	1,69	10
:	359	2,50	6
!	128	0,89	11
?	289	2,01	9
/	29	0,20	13
(	337	2,35	7
)	394	2,75	5
–	2.022	14,12	3
"	316	2,20	8
'	25	0,17	14
_	813	5,67	4
...	124	0,86	12

Os grafemas consonantais mais frequentes são: S (7,79%), R (6,59%), N (5,38%), D (5,14%), T (4,61%) e M (4,50%). Verifica-se que a frequência dos grafemas estangeiros (K, W e Y) é inexpressiva: 0,05%. Convém salientar que os dez grafemas mais frequentes cobrem 73,37% do total das ocorrências e que os cinco mais frequentes correspondem a 51,65%, isto é, mais do que metade de todas as ocorrências grafêmicas.

Retornando à Tabela 7, através das frequências percentuais e dos postos ocupados por cada grafema supragrafado, constata-se que (Ã) ocupa o primeiro posto com uma ocorrência de 28,91%. Contrariamente, o (Õ) ocupa um dos últimos postos com somente 2,85% de ocorrências. Outros grafemas acentuados a ocuparem altos postos na frequência são (É), (Á), (Í) e (Ó). Esses dados podem ser nitidamente visualisados no histograma do quadro seguinte.

Distribuição da frequência dos caracteres grafémicos supragrafados



Para avaliar a segunda hopótese (4.2.2), que pretende verificar a relação entre a distribuição das frequências e as amostras, calculou-se, inicialmente, a correlação simples entre as frequências dos caracteres grafêmicos e as amostras, duas a duas (Tabela 9).

Tabela 9 – Coeficientes de correlação simples entre as frequências dos caracteres grafêmicos e as amostras, duas a duas

AMOS - TRAS	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	TOTAL
I	-	.98	.97	.97	.98	.99	.97	.99	.94	.99	.99	.97	.99
II		-	.96	.97	.98	.98	.97	.98	.95	.98	.97	.99	.98
III			-	.99	.99	.97	.99	.98	.99	.98	.97	.97	.99
IV				-	.99	.97	.99	.99	.98	.99	.98	.97	.99
V					-	.98	.99	.99	.98	.99	.98	.97	.99
VI						-	.96	.99	.95	.99	.99	.97	.99
VII							-	.98	.98	.98	.97	.97	.98
VIII								-	.96	.99	.99	.97	.99
IX									-	.97	.96	.96	.97
X										-	.99	.97	.99
XI											-	.97	.99
XII												-	.98

Os altos coeficientes de correlação evidenciam que a tipologia textual não constitui variável interveniente na distribuição da frequência, o que queríamos provar. Isso, em outros termos, significa que a amostra global é suficientemente ampla, não necessitando de um corpus mais extenso. A significância desses dados fica abaixo do nível 0,01 visto que o valor crítico para esse nível é 0,48.

Se observarmos os coeficientes de correlação entre os caracteres intergrafêmicos e as diversas amostras (Tabela 10) verificaremos que o comportamento não é o mesmo. Os coeficientes, embora oscilem entre moderados e altos, nos levam a acreditar que, nesses caracteres, a tipologia textual constitui variável que altera os resultados. Para examinarmos a revelância dessa variável, calculamos o chi-quadrado das frequências. O chi-quadrado calculado foi de 4.141,07. Considerando que o valor crítico de

alfa, para o nível de significância de 0,05 é 0,532, conclui-se que existe diferença entre os textos no que concerne aos caracteres intergrafêmicos. Este achado pode ser mais claramente visualizado quando comparamos os números da ordem de frequência decrescente (postos) nas diversas amostras (Tabela 11).

Tabela 10 – Coeficiente de correlação simples entre as frequências dos caracteres intergrafêmicos e as amostras, duas a duas

AMOSTRAS	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	TOTAL
I	-	.94	.71	.93	.84	.91	.89	.96	.83	.95	.94	.92	.95
II		-	.80	.98	.91	.91	.85	.97	.81	.97	.97	.98	.98
III			-	.84	.95	.55	.86	.68	.92	.69	.80	.76	.86
IV				-	.94	.87	.88	.94	.83	.96	.97	.96	.98
V					-	.71	.89	.83	.86	.84	.89	.87	.94
VI						-	.67	.96	.62	.95	.89	.93	.87
VII							-	.81	.83	.80	.91	.81	.92
VIII								-	.71	.98	.95	.96	.94
IX									-	.72	.85	.83	.88
X										-	.96	.97	.95
XI											-	.97	.98
XII												-	.96
TOTAL													-

Tabela 11 – Ordem decrescente das frequências dos caracteres intergrafêmicos, em cada amostra

CARAC-TERES	AMOSTRAS											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
,	1	1	2	1	2	1	2	1	2	1	1	1
.	2	2	1	2	1	2	1	2	1	2	2	2
;	6	9	10	10	10	6	12	7	7	9	9	7
:	7	6	6	6	6	8	5	8	5	10	7	4
!	14	13	7	9	9	10	6	11	10	13	13	11
?	13	5	5	8	5	10	7	11	4	12	10	8
/	10	10	14	14	12	10	13	10	6	11	13	12
(	5	7	12	11	12	4	10	5	8	6	6	5
)	4	8	13	12	12	3	11	4	9	7	5	6
-	3	3	4	3	4	5	3	3	3	3	3	3
"	9	13	9	5	8	7	8	11	11	4	4	9
'	12	11	11	13	11	10	13	9	13	8	11	10
–	8	4	3	4	3	9	4	6	14	5	8	12
...	11	12	8	7	7	10	9	11	12	13	12	12

Enquanto as frequências dos caracteres (,), (.) e (–) ocupam praticamente os mesmos postos em todas as amostras, nota-se uma significativa discrepância (variação) nos postos de outros caracteres. Se repararmos a Tabela 12, frequência percentual, poder-se-á analisar melhor o comportamento de cada caracter intergrafêmico, em cada amostra.

Tabela 12 – Frequência percentual dos caracteres intergrafêmicos em cada amostra

CARA-	AMOSTRAS											
TERES	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
.	34,12	45,78	20,66	36,90	27,69	46,42	24,80	50,31	29,11	47,5	38,26	50,0
,	20,93	28,81	38,07	26,79	29,18	16,46	31,23	22,39	56,73	22,58	30,09	32,46
;	3,98	0,74	0,33	0,59	0,55	3,50	0,19	1,45	0,60	0,94	0,60	1,62
:	3,66	2,13	2,98	3,59	3,85	0,67	3,98	0,63	1,15	0,68	2,11	1,94
!	-	-	2,64	1,28	1,65	-	3,02	-	0,27	-	-	0,32
?	0,09	2,56	5,09	1,62	5,19	-	2,63	-	3,65	0,17	0,30	0,64
/	0,32	0,53	-	-	-	-	-	0,27	0,94	0,25	-	-
(	4,48	0,85	0,26	0,34	-	9,85	0,32	2,90	0,60	2,32	2,48	1,94
)	6,22	0,85	0,26	0,34	-	9,98	0,32	2,99	0,60	2,32	3,69	1,94
-	19,60	11,41	9,53	11,55	11,25	9,17	24,74	15,86	5,55	12,32	15,76	7,79
"	2,61	-	1,32	6,16	2,43	3,50	1,79	-	0,27	6,12	5,20	0,64
'	0,09	0,21	0,33	0,25	0,15	-	-	0,45	0,20	1,12	0,15	0,64
_	3,66	5,86	16,22	8,30	15,42	0,40	5,78	2,71	-	3,62	1,20	-
...	0,18	0,21	2,25	2,22	2,59	-	1,15	-	0,27	-	0,15	-

Finalmente, resta avaliar a última hipótese (4.2.3); ela se refere às frequências dos digramas. As tabelas 5, 6 e 7, que nos listam os digramas em ordem decrescente de ocorrência, confirmam a hipótese em referência. Vale a pena observar que o digrama DE ocupa o primeiro posto nas três posições. Verifica-se, outrossim, que os primeiros postos, tanto em posição inicial quanto medial, são ocupados pelos seguintes digramas: DE, CO, QU, SE, DO, ES, PA, UM, DA. Estes encontros são realmente os mais produtivos. Também chama a atenção o fato de que entre os encontros consonantais, listam-se os grupos PR, TR, GR, CR, CH, BR, FR, LH, PL,

CL, e que estes grupos ocupam postos semelhantes, um em relação ao outro, nas duas posições onde eles tem condições de aparecer: posição inicial e medial. Os únicos grafemas consonantais que aparecem em posição final são: L, S, M, R, N, Z.

5. Discussão dos resultados

Embora o objectivo imediato - objectivo satisfatoriamente atingido - não extrapole o plano meramente descritivo, existem diversas contribuições no plano teórico. Será, no entanto, o plano aplicativo que merecerá maior atenção em etapas subsequentes.

O objectivo imediato, vinculado à primeira e à terceira hipótese (3.2.1 e 3.2.3), foi atingido, como demonstra a análise estatística da distribuição de frequência. Tanto as frequências percentuais dos caracteres gráficos quanto dos digramas puderam ser devidamente ordenados.

A segunda hipótese - aquela relacionada com a estrutura sintático--semântico-estilística - fornecem dados suficientes para atingir o terceiro objectivo mediato. Chegou-se à conclusão de que o aspecto "estrutura textual" não constitui variável interveniente para alterar os gerais, no que se refere aos caracteres grafêmicos. Todos os coeficientes de correlação, calculados entre os diversos tipos de amostras, dois a dois, apresentaram-se muito fortes. As principais comparações feitas foram as seguintes:

a. Texto Científico X Texto Literário (amostras I e VII);
b. Discurso Dissertativo X Narrativo X Descritivo (amostras I, X e VIII);
c. Texto Completo X Parte de Texto (amostras I e VI);
d. Assuntos Diferentes: Linguística, filosofia, história e geografia (amostras I, XI, III e VIII);
e. Objectivo: informativo X didático (amostras X e VIII);

f. Texto Completo X Texto Constituído de minitextos(amostras I e IX);
g. Parte inicial, parte medial e parte final (amostras IV, V e III);
h. Prosa X Verso (amostras V e II);
i. Produção adulta X produção infantil (amostras V e XII);
j. Texto endereçado a adultos X texto endereçado a crianças (amostras V e VII).

No entanto, chama a atenção o fato de as correlações estabelecidas no plano dos caracteres intergrafêmicos não apresentarem comportamento similar. Foram verificadas correlações moderadas entre algumas amostras. A análise do chi-quadrado mostrou ser significativa a influência do tipo de amostra na distribuição da frequência desses caracteres. Faz-se, portanto, necessária uma melhor investigação desse campo. Talvez até se consiga encontrar, nos caracteres intergrafêmicos, determinantes capazes de discriminar diversas amostras entre si.

Por outro lado, os dados de frequência aqui levantados e computados deverão oportunuzar a comparação com dados de frequência de outros idiomas, dados já amplamente investigados e divulgados para o inglês, o francês e o alemão, entre outros.

Os resultados finais permitem partir para outras investigações e cálculos com os quais poderão ser alcançados os objectivos mediatos: contribuir na solição de problemas relacionados com editores de texto e com a disposição de caracteres em teclados de microcomputadores e oferecer subsídios ao ensino da linguagem, na área de alfabetização, principalmente no que concerne à gradação de material a ser apresentado ao aprendiz.

Uma das tarefas centrais será a maneira de aplicar os resultados da frequência, em conjunção com a facilidade de acessamento dos dedos às diferentes teclas, para um reordenamento dos teclados de máquinas eletrônicas de digitação. Além desse estudo de frequência, com o auxílio de um

fisiólogo, deverá ser avaliada a prontidão de reflexos dos diferentes dedos da mão e do esforço (trabalho) exigido aos mesmos dedos para impulsionarem teclas diferentes daquelas onde eles normalmente se posicionam. Os caracteres mais frequentes devem ocupar as teclas mais fáceis de serem acessadas; também devem ser tomadas en consideração as sequências grafêmicas mais frequentes. A cada tecla deve ser fornecido um número de ordem segundo a rapidez com que puderem ser acessados. Esta rapidez dependerá da capacidade de resposta de cada dedo a um estímulo enviado pelo cérebro e da distância que as teclas se encontram dos dedos escolhidos para acioná-las. No final desse estudo, procurar-se-á uma correlação positiva perfeita entre a frequência de ocorrência e a facilidade de acessamento. Os caracteres mais frequentes devem corresponder às teclas mais facilmente impulsionadas de modo a se obter o maior rendimento com o mínimo de custo.

Para a gradação do material de cartilhas de alfabetização, os dados levantados nesta pesquisa não poderão ser aplicados isoladamente. Far-se-á necessário um estudo de frequência vocabular, primeiro da língua portuguesa em geral e, depois dos vocábulos regionais. Estudos nesse sentido foram feitos, entre outros, por Sebastião (1983). Dos vocábulos selecionados privilegiar-se-ão aqueles que forem constituídos pelas letras e pelos encontros grafêmicos mais produtivos.

Ainda merecerão atenção, nesta seleção de material, os problemas oriundos da correspondência entre grafemas e fonemas, amplamente descritos e analisados por Silva (1981) e Lemle (1982: 41-60), e os originados pela interferência de línguas em conto, aspectos pesquisados por Tasca (1978) e Bisol (1986: 71-92).

Uma segunda tarefa consistirá em equacionar a gradação do material de alfabetização com a produtividade dos diversos caracteres grafêmicos, isoladamente ou em conjunção com os digramas mais frequentes. Deverão ser privilegiados os grafemas mais produtivos, não ignorando, no entanto, outras variáveis intervenientes como a facilidade (gráfica) de produzi-los.

O produto final desta pesquisa servirá de sugestão e não de imposição. Como existe a lei da inércia no mundo físico, o conservadorismo exagerado tem dificuldades em promover o avanço científico e tecnológico no mundo cultural. Considerando, no entanto, que esta sugestão se estriba em dados científicos, é de se supor que a tecnologia veja o alcance desta sugestão e dela faça uso num futuro não muito distante.

6. Conclusão

O objectivo operacional básico – levantar, em textos escritos no português do Brasil, a frequência dos caracteres gráficos e dos digramas em posição inicial, medial e final de palavras – foi satisfatoriamente atingido. As três hipóteses operacionais foram confirmadas.

A primeira – os caracteres gráficos em textos de língua portuguesa apresentaram percentagens de frequência diferentes – foi avaliada à luz da frequência percentual e dos postos ocupados por cada um dos caracteres grafêmicos.

A segunda – a tipologia das amostras não influi na distribuição dos caracteres grafêmicos – foi avaliada e confirmada pelo exame dos coeficientes de correlação simples entre a frequência dos caracteres grafêmicos e as amostras, duas a duas.

A terceira – certos digramas são mais produtivos do que outros – foi analisada e confirmada através das listas de digramas, apresentados em ordem decrescente de sua frequência.

O atingimento dos objectivos aplicativos constituirá uma etapa posterior, uma investigação e um estudo aditado à presente pesquisa. Nesse estudo deverá receber atenção especial o primeiro desses objectivos: Sugerir mudanças no actual teclado padrão QWERTY de microcomputadores e de outras máquinas eletrônicas de processamento de textos (editoração). Para alcançar esse objectivo, os resultados aqui apontados deverão ser cotejados

com levantamentos ergonométricos – reflexos dos diferentes dedos da mão e quantidade de trabalho exigido para acionar as diferentes teclas do teclado.

Por outro lado, deverão ser encontradas aplicações dos dadoos da frequência dos caracteres gráficos na elaboração cientifica de cartilhas de alfabetização.

Num terceiro momento, deverá ser feito um estudo comparativo dos dados aqui apresentados com os dados que outros pesquisadores já levantaram, na análise de outros idiomas.

Ao concluir o presente relatório convém que se saliente o aspecto descritivo que acompanhou a primeira fase desta pesquisa – mero levantamento de dados. A segunda fase – o que ainda está por ser feito – deverá privilegiar o aspecto aplicativo: teclados de microcomputadores e material de alfabetização.

BIBLIOGRAFIA

BARANOW, Ulf Gregor, Perspectivas na contribuição da linguística e de áreas afins a Ciência da Informação, *Ciência da Informação*, Brasília, CNPq/IBICIT, 12 (1): 23-35, 1983.

BISOL, Leda.,Interferência de uma segunda língua na aprendizagem da escrita, In: Tasca & Poersh, *Suportes Linguísticos para a alfabetização*. Porto Alegre, Sagra, 1986, pag. 71-92.

CHERRY, Colin, A comunicação humana, São Paulo, Cultrix, 1971.

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.

FEIGENBAUM, Edward and MacCOURDUCK, Pamela, *The fifth generation: artificial intelligence and Japan's computer challenge to the world*, New American Library. New York, 1984.

GRUPO EDUCAÇÃO E CULTURA, "O texto perfeito", *Software*, Rio, Rio Gráfica, 1984.

GRUPO EDUCAÇÃO E CULTURA, "Problemas no teclado", *Chips & bytes*, Rio, Rio Gráfica, 1984.

GUIRAUD, PIERRE, *Problémes et methodes de la statistique linguistique*, Dordrecht, D. Reibel Publishing Company, 1959.

HALLER, Johann, "Análise linguística e indexação automatica de textos", *Veritas*, Porto Alegre, PUCRS, 31 (123): 393-414, 1986.

HJELMSLEV, Louis, *Prolegômenos a uma teoria da linguagem*, São Paulo, Perspectiva, 1975.

HERDAN, Gustav, *The advanced theory of language as choice and chance*, Heidelberg, Springer-Verlag, 1966.

INTERNATIONAL SOLUTION, *The diplomat: installation manual*, Firth edition, Saratoga (Ca), International Solution, 1983.

LEMLE, Miriam, A tarefa da alfabetização: etapas e problemas do português. *Letras de Hoje*, Porto Alegre, PUCRS, (50): 41-60, dez. 1982.

LEPSCHY, Giulio C, *A linguística estrutural*, São Paulo, Perspectiva, 1971.

MALMBERG, Bertil, *As novas tendências da linguística*, São Paulo Nacional, 1971.

MAZZOCCO, Alexis (entrevista), "Opportunities for linguistics in the field of computers", *The linguistics reporter*, sep. 1979.

MILLER, George, *Language and communication*.New york, McGraw-Hill Company, 1951.

POERSCH, José Marcelino, *O linguista e a informática: relato* de uma contribuição, I Congresso Brasileiro de Linguística Aplicada: Resumos, Campinas, IEL, 1986.

POERSCH, José Marcelino, "Pode-se alfabetizar sem conhecimentos de linguística?", In: TASCA, Maria e POERSCH, José Marcelino. *Suportes linguísticos para a alfabetização*, Sagra, 1986.

POERSCH, José Marcelino, *Versão do Diplomata para a língua portuguesa; contribuição da linguística para a ciência da computação*, PUCRS, Centro de Pesquisas Linguísticas, 1986. Relatório de pesquisa.

SCHANK, Roger & CHILDERS, Peter, *The cognitive computer: on language, learning and artificial intelligence*, Menlo Park, Addison-Wesley Publishing Company, 1984.

SCHREIBER, Servan & Jacques, Jean (entrevista), "Informática e Informação", Veja, São Paulo, Editora Abril, (900): 3-6, 4 dez. 1985.

SILVA, Myriam Barbosa, *Leitura, ortografia e fonologia*, São Paulo, Ática, 1981.

TASCA, Maria, "A linguagem dos materiais de alfabetização", In: TASCA, Maria & POERSCH, José Marcelino, *Suportes linguísticos para a alfabetização*. Porto Alegre, Sagra, 1986.

VISÃO (autor não citado), "Inteligência artificial: o Brasil entra na corrida", *Visão*, pág. 34-38, 22 jan. 1986.

VOTRE, Sebastião Josué, "Por uma linguística Aplicada à alfabetização", *Letras de Hoje*, Porto Alegre, PUCRS 13 (42): 20-34, dez. 1980.

VOTRE, Sebastião Josué, *Um léxico para cartilha*, Rio de Janeiro, Universidade Gama Filho MEC/INEP, 1983.

ZIPF, G. K., *human behavior and the principle of least effort*, Cambridge (Mass.), Addison-Wesley Publishing Company, 1949.