Fonologia: Contribuições para a Linguística e para a Computação

(Phonology: Contributions to Linguistics and Computational Science)

Thaïs Cristófaro Silva[1]

[1]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)

thaiscristofarosilva@ufmg.br

Abstract:. This paper discusses how the discipline of Linguistics could cooperate with Computational studies. The focus of the discussion is on phonological studies. It is intended to address how we could help computers, and other equipments, to interact with their users using speech. It is argued that choosing a given theoretical perspective is crucial to formulate specific tools that will contribute towards human-machines interactions. It is claimed that the tools must be formulated within a multi-disciplinary perspective.

Keywords: Computational Linguistics; Phonology; Database.

Resumo: Este trabalho tem por objetivo discutir as contribuições da Linguística para a Computação, com ênfase no domínio da Fonologia. Pretende-se, portanto, contribuir com o debate de como poderemos ajudar o computador, e outros equipamentos, a interagir com seus usuários utilizando a sonoridade. Argumenta-se que o enfoque teórico é crucial para a formulação de ferramentas específicas, as quais devem ser construídas multidisciplinarmente.

Palavras-chave: Linguística Computacional; Fonologia; Banco de Dados.

Introdução

Este trabalho tem por objetivo discutir as contribuições da Linguística para a Computação, com ênfase no domínio da Fonologia.[1] A Fonologia é a disciplina da Linguística que busca compreender e explicar a organização gramatical da sonoridade. Para que seja possível formular um equipamento que interaja com as pessoas através da fala devemos compreender e explicar como a sonoridade se organiza. Por essa razão a Fonologia pode contribuir com o debate de como poderemos ajudar o computador, e outros equipamentos, a interagir com seus usuários utilizando a sonoridade. Espera-se que ao avaliar a interface entre a Linguística e a Computação este trabalho contribua com o debate teórico da Linguística e ao mesmo tempo ofereça instrumentos importantes para a implementação de recursos da linguagem utilizando o computador.

Este trabalho tem a seguinte organização. A primeira seção discute duas abordagens teóricas opostas. Busca-se indicar que um determinado enfoque teórico adotado é crucial para a implementação de recursos tecnológicos de interface entre a Linguística e a Computação. A segunda seção apresenta o Projeto ASPA (Avaliação Sonora do Português Atual), que é uma ferramenta de busca fonológica gerenciada em banco de dados. Esta seção também ilustra alguns casos de utilização do banco de dados do projeto ASPA e aponta para caminhos futuros de investigação. A terceira seção apresenta o projeto e-Labore (Laboratório Eletrônico

[1] A autora agradece ao apoio do CNPq através de Bolsa de Produtividade em Pesquisa, Processo 304076/2008-2 e o apoio ao Projeto de Pesquisa CS, Processo 401099/2009-1. A autora agradece também à FAPEMIG pelo apoio através do Programa Pesquisador Mineiro PPM-IV, Processo 16415.

de Oralidade e Escrita) que consiste de um banco de dados de produções textuais infantis. Esta seção também ilustra a possível utilização do banco de dados. A quarta seção avalia a relação entre teorias linguísticas e recursos tecnológicos, indicando possíveis caminhos a serem trilhados no futuro. Finalmente, a quinta seção apresenta a conclusão e é seguida das referências bibliográficas.

Abordagens teóricas

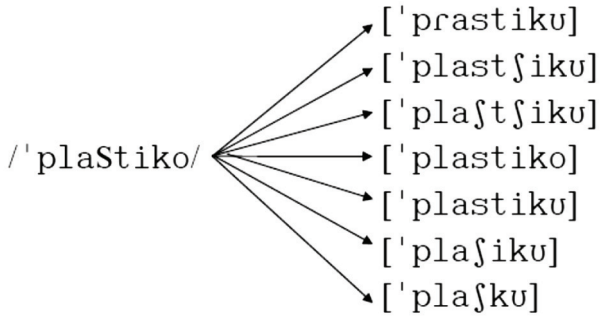
Esta seção discute duas abordagens teóricas que se opõem com relação à organização do conhecimento linguístico. Os aspectos centrais dessas abordagens teóricas serão apresentados, bem como as consequências de tais abordagens para a formulação de recursos tecnológicos de apoio à análise linguística. Busca-se indicar que o enfoque teórico adotado é crucial para a implementação de recursos tecnológicos de interface entre a Linguística e a Computação.

As várias teorias linguísticas discordam entre si às vezes substancialmente e às vezes pontualmente. A discordância é salutar por implementar o debate e avançar as concepções teóricas. Embora haja muita discordância entre as várias abordagens linguísticas há consenso entre elas de que a linguagem tem caráter abstrato. O debate teórico centra-se, sobretudo, em relação a como se dá a organização do conhecimento linguístico abstrato da linguagem. Neste trabalho a discussão desse tema se centrará no conhecimento fonológico.

A abordagem tradicional, que de alguma maneira consiste da base teórica inicial da linguística, sugere que o conhecimento linguístico tenha alto grau de abstração e que as representações linguísticas sejam simples (SAUSSURE, 1916; CHOMSKY; HALLE, 1968). Assim, informações redundantes são excluídas das representações e serão incorporadas através de gerenciamento complexo. As teorias gerativas, de maneira geral, representam essa perspectiva.

Considere a figura 1, que ilustra a relação entre a representação fonológica da palavra *plástico* e algumas de suas várias pronúncias. Como é tradicionalmente assumido, a representação fonológica é apresentada entre barras transversais: /'plaStiko/. As várias representações fonéticas são apresentadas entre colchetes.

Figura 1. Representação fonológica e fonética da palavra *plástico*



A representação fonológica é assumida ser simples e exclui informações redundantes. O processamento para as várias formas fonéticas se dá por gerenciamento complexo, que pode ser processual como assumido nos modelos gerativos clássicos (KENSTOWICZ,

1994; GOLDSMITH, 1990), ou por restrições, como assumido na Teoria da Otimalidade (KAGER, 1999). Na abordagem tradicional as representações são simples e o processamento é complexo. A abstração é gerenciada por princípios da Gramática Universal. Uma vez inferido o mecanismo que gerencia a Gramática não há necessidade de *corpora* para análise. Ou seja, a análise sendo evidenciada permite o gerenciamento de qualquer *corpora*. Assim, efeitos probabilísticos são descartados nos modelos tradicionais.

Uma abordagem alternativa que tem sido desenvolvida nos últimos anos sugere a multirrepresentacionalidade (CRISTÓFARO SILVA; GOMES, 2007). A Figura 2 apresenta a representação de um conjunto de exemplares para a palavra *plástico*. Esse tipo de representação é proposto pela Teoria de Exemplares (JOHNSON, 1997; PIERREHUMBERT, 2001).

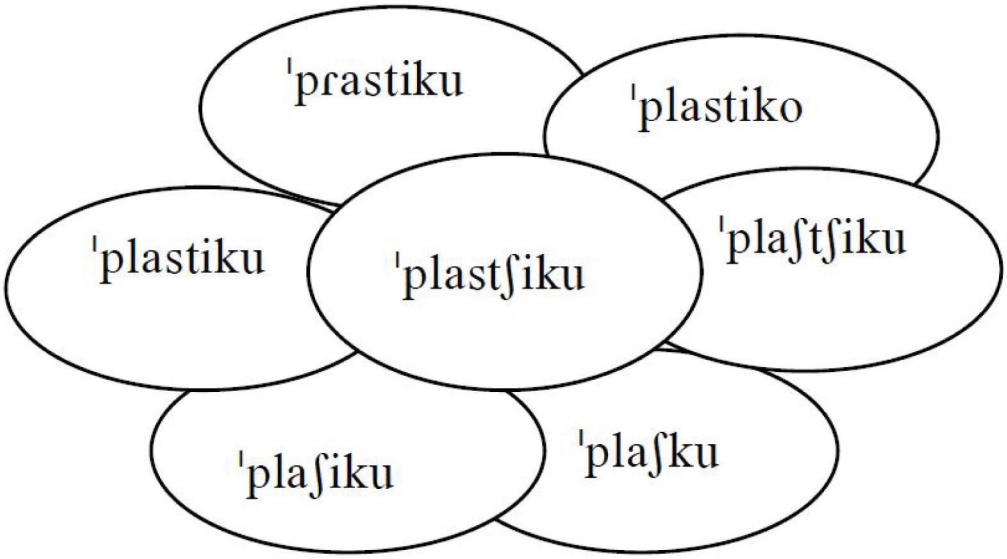


Figura 2. Representação dos exemplares da palavra *plástico*

Os exemplares representam instâncias de uso, compreendendo a produção e a percepção da experiência linguística do falante. Os exemplares agregam informações linguísticas contextuais e também informações tradicionalmente compreendidas como não linguísticas como, por exemplo, informações sociofonéticas. Por essa razão postula-se a multimodalidade do conhecimento linguístico na Teoria de Exemplares. O gerenciamento dos exemplares se dá probabilisticamente. A palavra é o lócus representacional. Assim, efeitos de frequência são cruciais para a organização do conhecimento linguístico (BYBEE; HOPPER, 2001; BOD; HAY; JANNEDY, 2003). Nesta abordagem as representações são complexas e o processamento é simples.

A Teoria de Exemplares é o modelo representacional assumido pela Fonologia de Uso (BYBEE, 2001, 2010). Por incorporar efeitos de frequência, a Teoria de Exemplares sugere a utilização de *corpora*. *Corpora* demandam o gerenciamento de grande volume de dados e, por esta razão, é comum que linguistas e profissionais da área de tecnologia e ciências da fala atuem em conjunto para a organização de *corpora*. Em *corpora* específicos os efeitos de frequência podem ser observados e avaliados.

A discussão apresentada nesta seção indica que a concepção teórica adotada contribui para a implementação de metodologias específicas. A Teoria de Exemplares e a Fonologia de Uso sugerem a pertinência, a adequação e a relevância do uso de *corpora*. Foi nesse contexto teórico que os projetos ASPA (Avaliação Sonora do Português Atual) e e-Labore (Laboratório Eletrônico de Oralidade e Escrita) foram formulados. Cada um desses projetos será apresentado nas próximas seções.

Projeto ASPA: Avaliação Sonora do Português Atual

Esta seção apresenta o Projeto ASPA: Avaliação Sonora do Português Atual. O ASPA é um empreendimento conjunto entre pesquisadores que atuam em áreas diversas do conhecimento e que necessitam de um conhecimento sólido da organização sonora do português contemporâneo. O entrelace maior entre esses pesquisadores é a concepção teórica de que o conhecimento linguístico é organizado probabilisticamente. Informações sobre o ASPA são disponibilizadas em www.projetoaspa.org. Informações sobre a formulação inicial do ASPA podem ser obtidas em Cristófaro Silva e Almeida (2005) e Almeida (2005). Os resultados de buscas específicas no banco de dados do ASPA oferecem subsídios para pesquisas em diversas áreas do conhecimento, dentre estas: teorias linguísticas, teorias fonéticas e fonológicas, ensino de fonética e fonologia, linguística de *corpus*, linguística aplicada à educação, organização de banco de dados, linguística computacional e formulação de *software*.

A lista de palavras que foi utilizada pelo projeto ASPA é composta por um total de 607.392 palavras diferentes que totalizam 228.766.402 de palavras em geral. Os dados de origem do Projeto ASPA são provenientes de uma lista de palavras fornecida, em 2004, pelo Projeto DIRECT-PUC-SP: http://www2.lael.pucsp.br/corpora/. Os dados ortográficos fornecidos foram convertidos para um código específico de cadastro da sonoridade, o LETRASOM (CRISTÓFARO SILVA; ALMEIDA, 2005; ALMEIDA, 2005). Além da conversão automática foi necessário o cadastro de informações específicas como, por exemplo, a categoria morfológica, bem como foi realizada uma avaliação geral dos dados visando a excluir siglas, dados com desvio da ortografia vigente e palavras de outras línguas diferentes do português.

Visando à operacionalidade do trabalho de transcrição, foram cadastradas no banco de dados do Projeto ASPA palavras cuja frequência de ocorrência no *corpus* fosse maior ou igual a 6. Assim, das 607.392 palavras do *corpus* original foram transcritas 199.864. Portanto, o número de tipos considerados para a transcrição pelo LETRASOM foi 199.864. Tais tipos totalizaram 10.739.395 ocorrências. O banco de dados permite aos usuários fazerem observações quanto aos dados cadastrados. Assim, pode haver alteração nesses números caso haja sugestão de algum usuário em que seja pertinente a adequação dos dados do *corpus*. Ao efetuar qualquer busca no banco de dados do Projeto ASPA, o usuário terá como resultado o padrão sonoro buscado, bem como informações sobre a frequência de tipo e sobre a frequência de ocorrência do padrão buscado. Segundo Bybee (1985, 2001), o armazenamento e o processamento dos itens lexicais estão sujeitos tanto a efeitos de frequência de tipo, quanto a efeitos de frequência de ocorrência.

A frequência de tipo (*type frequency*) corresponde à frequência de um padrão específico no léxico (ou dicionário). Uma palavra, por exemplo, pode ser considerada um tipo específico

em um determinado *corpus*. A palavra *vida* é um tipo dentro do léxico do português brasileiro. A sílaba *vi* é também um tipo da Gramática Fonológica do português. Assim, na sentença "A vida é para ser vivida com vigor.", observamos que a frequência de tipo da palavra *vida* é 1 e a frequência de tipo da sílaba *vi* é 4: *A vi*^{1}*da é para ser vi*^{2}*vi*^{3}*da com vi*^{4}*gor*., uma vez que a sílaba *vi* ocorre 4 vezes nesta sentença.

Por outro lado, a frequência de ocorrência (*token frequency*) corresponde ao número de vezes que um determinado elemento ocorre em um *corpus*. Por exemplo, se buscarmos a palavra *vida* em um determinado *corpus* do português brasileiro e encontrarmos o índice 112.365, podemos afirmar que a frequência de ocorrência da palavra *vida* é de 112.365. Pode-se buscar a frequência de ocorrência em vários níveis como, por exemplo, uma palavra, um morfema, um padrão silábico, um segmento, etc. Por exemplo, se buscarmos em um *corpus* do português brasileiro o número de palavras que têm a sílaba *vi* e encontrarmos o índice de frequência de ocorrência de 26.481, podemos afirmar que a sílaba *vi* foi encontrada 26.481 vezes no *corpus* examinado. Por outro lado, se encontrarmos o índice de frequência de ocorrência de 45.224 para a sílaba *da* podemos afirmar que a sílaba *da* foi encontrada 45.224 vezes no *corpus* examinado. Isso nos leva a concluir que, na língua em questão, a sílaba *da* é mais produtiva do que a sílaba *vi*, uma vez que a sílaba *da* apresenta frequência de tipo mais alta que a sílaba *vi*. Considere a Figura 3, que ilustra a página inicial do buscador do ASPA.[2]

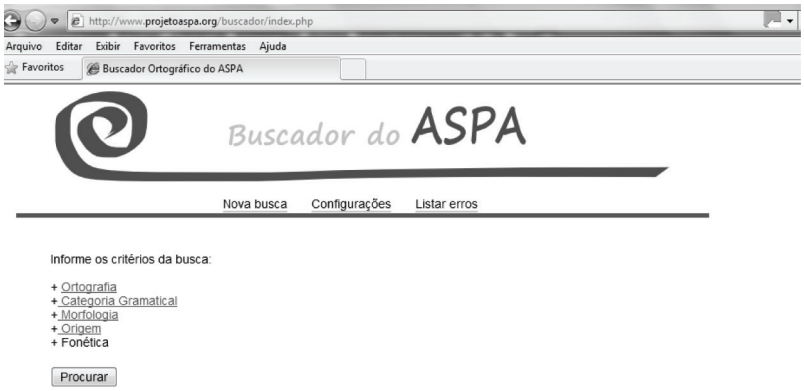


Figura 3. Página inicial do buscador do ASPA

A Figura 3 indica que a busca pode ser realizada em várias categorias: ortográfica, categoria gramatical, morfológica, origem da palavra e fonética. As buscas podem também combinar categorias, como exemplificado na Figura 4.

[2] Os dados apresentados neste artigo representam o estágio atual do buscador do ASPA. Encontra-se em curso uma revisão do buscador que deverá ser lançada em 2011 e que deverá ter interface mais objetiva para as buscas a serem realizadas pelos usuários.

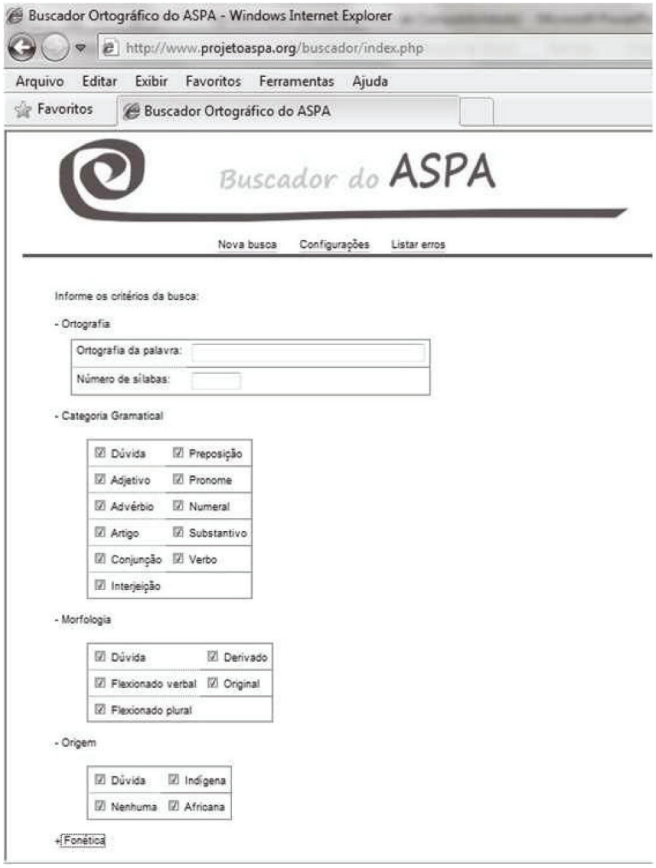


Figura 4. Desdobramentos da página do buscador do ASPA

Ao realizar uma busca, o usuário terá acesso a um arquivo texto que contém a lista das palavras com o padrão de busca realizado, bem como informações sobre a frequência de tipo e a frequência de ocorrência para o padrão. O Quadro 1 ilustra o resultado do buscador do ASPA para o padrão de (sibilante+rótico).

Quadro 1. Resultado de buscas para sequências (sibilante+rótico)

RESULTADO (sibilante+rótico)					
TIPOS: 70					
TOKENS: 43.943					
Índice	Frequência	Ortografia			
1125	21728	israel	98104	28	disritmia
3213	7458	israelense	99636	27	desregular
3835	6063	israelenses	99670	27	disraeli
8142	2331	desrespeito	104809	24	desregula
13834	1099	desregulamentação	104810	24	desregulamentou
14750	999	israelita	106778	23	desregulamentados
19272	658	desrespeitar	108894	22	desrecalque
28074	349	desrespeita	111084	21	desregrado
28790	335	desrespeitando	113456	20	desreguladas
31109	292	desrespeitado	115959	19	desregulamentada
32223	274	desrespeitou	124692	16	desrespeitava
39887	185	desrespeitam	128161	15	desrespeitasse
45647	143	desrespeitada	128162	15	desrespeitosos
47971	130	desrespeitosa	132003	14	desrepressão
48765	126	desrespeitaram	132004	14	desrespeitaria

49789	121	desrespeitados	136217	13	desrespeitassem
52271	110	israelitas	140911	12	desregulagem
53421	105	desregulamentar	140912	12	desregulamentações
57295	91	desrespeitarem	140913	12	desregulamentando
59848	83	desrespeitoso	146207	11	desregulamenta
60563	81	desregulação	152326	10	desrespeitavam
61653	78	desregulado	152327	10	desrespeitosamente
66699	66	desregramento	159325	9	desrealização
66825	66	israelândia	159326	9	desregulamentadas
69173	61	desregrada	159327	9	desrespeitei
70843	58	desrazão	167433	8	desregrados
70844	58	desrespeitos	167434	8	desregramentos
75089	51	desregulados	167435	8	desregulagens
77232	48	desregulamentado	167436	8	desregulamentaram
78694	46	desrespeitadas	177293	7	desregulando
83722	40	desrespeitem	186298	6	antiisraelense
89940	34	desregulada	189521	6	desrealiza
91129	33	desratização	189522	6	desrespeitamos
95148	30	desrespeitosas	189677	6	disruptivas
98086	28	desrespeite	192682	6	israelista

O Quadro 1 indica que foram encontradas 70 palavras que contêm o padrão de sibilante em final de sílaba sendo seguida de um rótico. Assim, a frequência de tipo do padrão (sibilante+rótico) é 70. A frequência de ocorrência desse padrão é de 43.943. O Quadro 1 ilustra os resultados conforme exportados pelo buscador do ASPA. A visualização em tela da busca desse mesmo padrão é apresentada na Figura 5.

A Figura 5 ilustra a busca solicitada para o padrão de sílabas terminadas em consoante sibilante (som associado ao som de S), sendo seguidas de róticos (sons associados ao som de R).

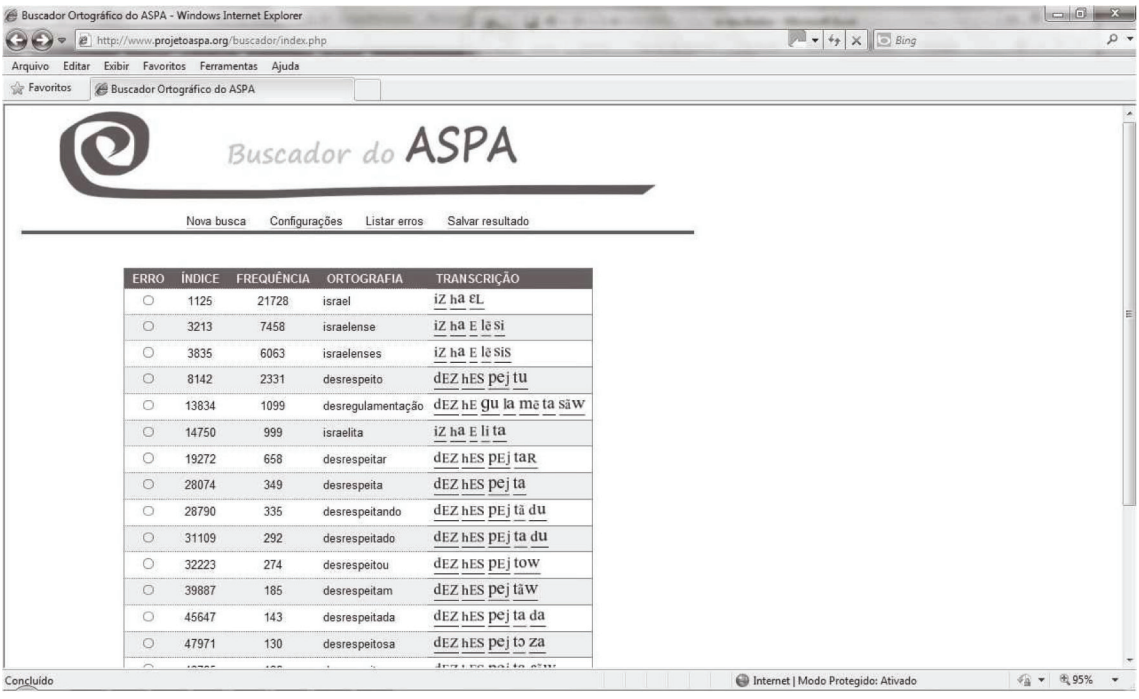


Figura 5. Informação em tela de busca realizada para sequência de sibilante e rótico

Além das informações exportadas para o arquivo texto de resultados, a primeira coluna dos resultados tem um botão com o qual o usuário poderá indicar algum erro de cadastro. A avaliação quanto à adequação do erro é realizada pelo gerenciador do banco de dados e, se pertinente, é feita a alteração no banco de dados.

Qual a relevância de se conhecer padrões sonoros específicos? Consideremos o caso discutido acima. No português, sempre que uma sílaba terminada em consoante é seguida de um som rótico, este será manifestado como um R-forte. O R-forte se opõe ao r-fraco em português, em posição intervocálica, como, por exemplo, nas palavras *caro* e *carro*. O r-fraco é sistematicamente pronunciado como um tepe alveolar [ɾ] em todas as variedades do português. O R-forte, por outro lado, apresenta ampla variação dialetal. Assim, no caso de sequências de (sibilante+rótico), sabemos que o R-forte representará o rótico. Os dados do buscador do ASPA indicam que o padrão (sibilante+rótico) é pouco frequente, englobando 70 palavras apenas. Podemos observar também que várias das palavras na categoria de (sibilante+rótico) apresentam o prefixo des-, ou seja, englobam tipicamente palavras morfologicamente complexas. Padrões pouco frequentes podem ser sujeitos a alterações para se ajustarem em padrões recorrentes da língua.

Consideremos agora um caso de mudança sonora envolvendo sequências de (lateral+rótico) em português (CRISTÓFARO SILVA; OLIVEIRA, 2002). Em variedades linguísticas em que ocorreu a vocalização da lateral, os autores observaram que, ao invés do R-forte, a população jovem estava fazendo uso do r-fraco. Assim, uma palavra como *guelra*, em que, tipicamente, o rótico ocorre como o R-forte, passa a ocorrer com um r-fraco: gue[wɾ]a. A pronúncia gue[wɾ]a não é esperada porque a fonologia do português prevê que após consoantes o rótico será sempre um R-forte (como previsto para as sequências de (sibilante+róticos) discutidas anteriormente). Contudo, houve a vocalização da lateral, passando a ocorrer não mais uma consoante e sim um glide posterior [w]. Quando glides posteriores são seguidos de róticos, em português, o rótico se manifestará como um tepe: *áurea*, *Europa*, *couro*, etc. Tendo conhecimento desses fatos, podemos explicar por que a pronúncia gue[wɾ]a passa a ocorrer no português, mesmo que em princípio seja não esperada. A pronúncia gue[wɾ]a decorre do fato que a vocalização da lateral cria uma situação em que um glide posterior é seguido de um rótico. Antes da lateral ser vocalizada, o rótico que a seguia era sistematicamente um R-forte. Contudo, considerando-se que o número de palavras com o padrão (lateral+rótico) é bastante pequeno (14 itens no ASPA), houve a inovação e o r-fraco passa a ocorrer. No padrão (glide posterior+rótico) ocorre, sistematicamente, o r-fraco. Ajusta-se então um padrão menos frequente — de (lateral+rótico) — a um padrão mais frequente (glide posterior+rótico).

A análise discutida acima nos mostra que, além de conhecermos os fatos do percurso da mudança linguística, podemos explicar por que o padrão inovador passa a ocorrer: em decorrência de efeitos de frequência. Sabemos também que é a população de faixa etária mais jovem que faz uso do padrão inovador (CRISTÓFARO SILVA; OLIVEIRA, 2002). Assim, podemos sugerir que ferramentas que visem à interação homem-máquina através da sonoridade incluam esse tipo de informação em sua implementação. Ao selecionar a faixa etária, o usuário seria direcionado para o padrão inovador enquanto falantes de faixa etária idosa teriam acesso à pronúncia tradicional. A vantagem, neste caso, é que usuário teria acesso à informação compatível com sua faixa etária. Se tivermos estudos relativos a outras informações fonológicas, estas podem ser incorporadas em ferramentas de interação

homem-máquina. Por exemplo, tendências observadas nas falas femininas e masculinas, ou tendências observadas quanto à região geográfica, etc. Tais informações podem oferecer ao usuário a possibilidade de ajustar a fala com que irá interagir.

Esta seção apresentou o Projeto ASPA, dando ênfase à apresentação do buscador do ASPA e de sua utilização na pesquisa linguística. Buscou-se indicar a relevância do conhecimento probabilístico da linguagem e da contribuição de análises baseadas em *corpora* em para a implementação de ferramentas tecnológicas que façam uso da sonoridade na interação homem-máquina. A próxima seção apresenta o Projeto e-Labore.

Projeto e-Labore: Laboratório Eletrônico de Oralidade e Escrita

Esta seção apresenta o Projeto e-Labore: Laboratório Eletrônico de Oralidade e Escrita. O objetivo central do Projeto e-Labore é o de coletar, cadastrar e disponibilizar para a comunidade científica um banco de dados de material escrito por crianças de 6 a 12 anos, residentes na cidade de Belo Horizonte (MG). Informações específicas sobre o projeto podem ser consultadas em Cristófaro Silva et al (2006, 2007) e encontram-se também disponibilizadas em www.projetoaspa.org/elabore. De maneira análoga ao Projeto ASPA, o Projeto e-Labore assume a relevância dos estudos de *corpora* na análise linguística, entendendo ser a palavra o lócus da representação lexical e argumenta pela organização probabilística do conhecimento linguístico.

O *corpus* do projeto e-Labore permite o mapeamento do vocabulário infantil do português brasileiro contemporâneo que pode oferecer contribuições para a investigação de teorias de aquisição da linguagem bem como pode contribuir com os debates a respeito da interação entre a linguagem adulta e infantil em um contexto de mudança linguística e evolução da linguagem.

Foram coletadas, digitalizadas e digitadas 7.892 produções textuais de crianças de 6 a 12 anos, estudantes de 1ª a 6ª Série do Ensino Fundamental, totalizando 270 turmas em 36 escolas diferentes. O *corpus* do Projeto e-Labore conta com um total de 821.731 palavras (frequência de ocorrência) sendo 22.610 palavras individuais (frequência de tipo). A Tabela 1, que segue, lista as palavras mais frequentes do *corpus* do Projeto e-Labore.

Tabela 1: Lista de frequência das palavras do corpus e-Labore

	Palavras			Substantivos			Verbos		
	Palav.	**Freq.**	**%**	**Subst.**	**Freq.**	**%**	**Verb.**	**Freq.**	**%**
1	e	9.622	4,64	natal	1.137	0,55	é	2.995	1,44
2	que	5.997	2,89	dia	1.108	0,53	foi	1.218	0,59
3	o	5.975	2,88	pessoas	993	0,48	tem	1.006	0,48
4	a	5.920	2,86	casa	747	0,36	era	833	0,40
5	de	5.528	2,66	mãe	589	0,28	estava	694	0,34
6	um	3.330	1,60	escola	510	0,25	ser	584	0,28
7	eu	3.307	1,60	ano	500	0,24	são	492	0,24
8	para	3.206	1,54	gente	496	0,24	tinha	491	0,24
9	não	3.058	1,47	mundo	446	0,22	está	481	0,23
10	é	2.993	1,44	projeto	384	0,19	vai	459	0,22
11	uma	2.268	1,09	crianças	381	0,18	fazer	429	0,21
12	com	1.969	0,95	família	327	0,16	ter	408	0,20
13	os	1.802	0,87	pai	323	0,16	vou	391	0,19
14	no	1.739	0,84	anos	323	0,16	pode	350	0,17
15	na	1.667	0,80	violência	315	0,15	acho	327	0,16

A Tabela 1 indica que várias das palavras mais frequentes representam monossílabos, incluindo palavras funcionais (itens de 1 a 15 na Tabela), bem como casos de morfologia irregular como, por exemplo, formas verbais flexionadas dos verbos *ser, estar, ser* e *ir*. Generalizações podem, portanto, ser feitas com relação à utilização da linguagem por crianças de diferentes faixas etárias.

O conhecimento da linguagem infantil pode também oferecer contribuições para a investigação dos problemas atestados no processo de apropriação da linguagem escrita pelas crianças em idade escolar. Encontra-se em fase de cadastramento os desvios ortográficos atestados na escrita infantil.

Os desvios ortográficos são comuns na fase de apropriação da linguagem escrita, e persistem, por vezes, após o período escolar. A interferência da oralidade na escrita permite que os desvios ortográficos sejam compreendidos e explicados adequadamente (ALVARENGA et al., 1989). A relação entre a oralidade e a escrita é mediada pelos estudos da sonoridade que englobam a fonética e a fonologia.

A identificação dos desvios ortográficos foi gerenciada automaticamente. Isso porque, ao serem digitadas, as redações tiveram marcações específicas. Por exemplo, o desvio ortográfico é indicado entre chaves {socego} sendo imediatamente seguido pela forma ortográfica convencional entre colchetes: [sossego]. Portanto, uma busca automatizada nos oferece dados importantes sobre o acervo do Projeto ASPA. Foram catalogadas 85.659 palavras grafadas com algum tipo de desvio ortográfico, representando 10,5% das palavras do *corpus* (821.731/85.659). Esses resultados expressam a ótima notícia de que os estudantes, de fato, têm alto índice de acerto no texto escrito, ou seja, os estudantes apresentam em torno de 90% das palavras grafadas de acordo com as normas ortográficas vigentes.

Por outro lado, se considerarmos as palavras que apresentaram a grafia com desvio ortográfico observamos que em torno de 34% das palavras foram grafadas equivocadamente. Ou seja, das 22.610 palavras distintas catalogadas no *corpus* do Projeto e-Labore 7.633 apresentaram algum tipo de desvio ortográfico.

Há uma questão paradoxal nestes dados: as crianças escrevem corretamente a maioria das palavras (90% de acerto), mas há um conjunto grande de palavras que apresenta desvios ortográficos (34% das palavras foram grafadas inadequadamente). De fato, esses dados revelam que escrever com algum tipo de desvio ortográfico no período de apropriação da escrita é, de alguma maneira, esperado. De fato, somente 2,35% das produções textuais catalogadas no *corpus* do Projeto e-Labore não apresentaram erros ortográficos (184/7.817). Cabe-nos fazer a seguinte pergunta: quantos erros ortográficos em um texto seriam considerados como um índice razoável do percurso de apropriação da escrita? Considere a Figura 6, que ilustra o número de desvios ortográficos atestados nas produções textuais do *corpus* do e-Labore.

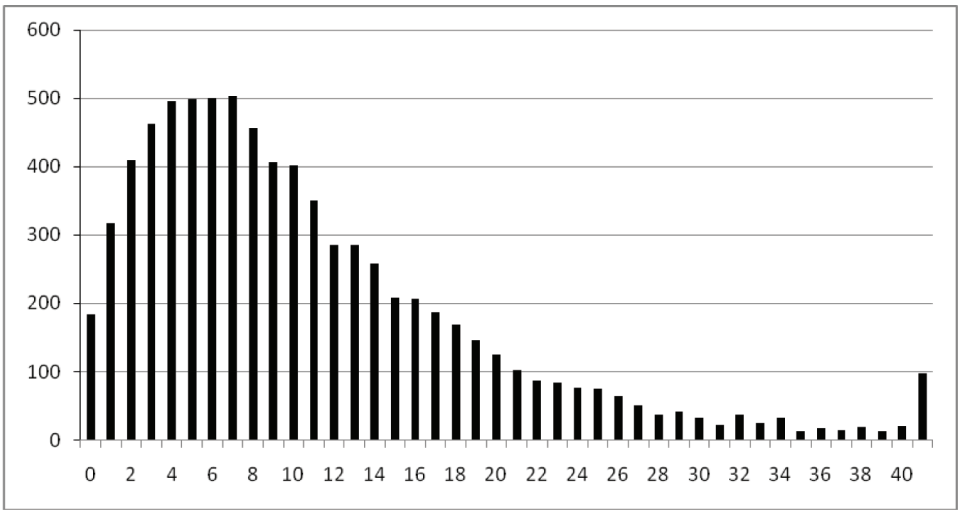


Figura 6. Quantidade de desvios ortográficos

O eixo das abscissas indica o número de erros atestados em uma única redação. O eixo das ordenadas indica o número de redações que apresentou a quantidade de erros indicada no eixo das abscissas. Como mencionado anteriormente, 184 produções textuais não apresentaram erros ortográficos. Essa informação aparece na primeira barra vertical do gráfico da Figura 6. Pode-se observar que a maioria das produções textuais apresenta em torno de 1 a 16 erros ortográficos. Este é o caso para 77% das produções textuais. Ou seja, um total de 6046 produções textuais apresentou de 1 a 16 erros ortográficos. Esses resultados indicam que apresentar erros ortográficos é relativamente comum no processo de apropriação da linguagem escrita, sendo que, se a quantidade de erros atestados por redação estiver entre 1 e 16 é, de alguma maneira, esperado. Acima de 17 erros em uma mesma redação pode ser considerado um padrão que começa a sair do comportamento geral atestado.

Os resultados apresentados nesta seção indicam que informações extraídas de um banco de dados de produções textuais infantis podem oferecer dados importantes sobre o desenvolvimento da apropriação da linguagem escrita e do uso do português escrito. Tais resultados, bem como outros de natureza semelhante, podem contribuir com a formulação de propostas que visem a melhorar o ensino do português escrito. Pode-se também buscar informações que instrumentalizem as professoras quanto às produções textuais infantis.

Esta seção apresentou o Projeto e-Labore dando ênfase para a avaliação dos desvios ortográficos atestados nas produções textuais de crianças de 6 a 12 anos. Buscou-se indicar a relevância do conhecimento probabilístico da linguagem e da contribuição de análises baseadas em *corpora* para a implementação de ferramentas tecnológicas que façam uso da sonoridade na interação homem-máquina.

Teorias linguísticas e recursos tecnológicos

Esta seção avalia a relação entre teorias linguísticas e recursos tecnológicos indicando possíveis caminhos a serem trilhados no futuro. O primeiro aspecto a ser destacado é o caráter dinâmico dos bancos de dados que organizam informações sobre a linguagem. A dinamicidade decorre do fato de o banco de dados ser ampliado ou o mesmo poder sofrer alterações decorrentes de avaliações metodológicas específicas. Um lado interessante da abordagem dinâmica do gerenciamento de *corpora* é a relação direta com as línguas naturais, que são essencialmente dinâmicas. Os modelos teóricos que apresentamos na parte inicial deste trabalho — Bybee (2001), Johnson e Mullenix (1997) e Pierrehumbert (2001) — acomodam a perspectiva dinâmica do gerenciamento de *corpora*. Portanto, a implementação de recursos tecnológicos específicos depende do enfoque teórico adotado.

Avanços teóricos podem implicar a reorganização de aspectos metodológicos. Tal reorganização permite-nos expressar o caráter dinâmico da linguagem. Como mencionado na segunda seção deste trabalho, há um projeto em curso que tem por objetivo gerenciar de maneira mais eficiente o buscador do Projeto ASPA. O gerenciamento mais eficiente do ASPA implica alterações metodológicas específicas. Esse aspecto é não apenas esperado, mas também permite expressar a dinamicidade das línguas naturais.

O Projeto e-Labore também poderá enfrentar desafios metodológicos quando, por exemplo, se der a ampliação do acervo. Assim, além de dados coletados na cidade de Belo Horizonte, será possível ampliar a coleta de natureza análoga ao Projeto e-Labore em outras localidades.

Finalmente, os bancos de dados do ASPA e do e-Labore demonstram esforços conjuntos entre profissionais de diversas áreas do conhecimento para a construção de recursos que possam contribuir com uma compreensão mais ampla da linguagem. As perguntas teóricas formuladas pelos diversos profissionais se interceptam e oferecem a possibilidade de uma avaliação da linguagem em várias perspectivas.

Com relação à interação homem-máquina espero que este artigo contribua para demonstrar a relevância de empreendimentos multidisciplinares que possam contribuir com o desenvolvimento de recursos tecnológicos específicos. A experiência dos bancos de dados do ASPA e do e-Labore são tentativas nesse sentido.

Conclusão

Este trabalho teve por objetivo discutir as contribuições da Linguística para a Computação, com ênfase no domínio da Fonologia. Sugere-se que, para que seja possível formular equipamentos que interajam com as pessoas através da fala, devemos compreender e explicar como a sonoridade se organiza. A compreensão da sonoridade deve estar inserida em perspectiva multidisciplinar, que envolva participantes de várias áreas do

conhecimento. Para defender essa proposta foram apresentados dois estudos de casos relacionados com a construção de banco de dados: 1) o Projeto ASPA (Avaliação Sonora do Português Atual), que é uma ferramenta de busca fonológica, e 2) o projeto e-Labore (Laboratório Eletrônico de Oralidade e Escrita), que consiste de um banco de dados de produções textuais infantis. Os parâmetros adotados na formulação de cada um dos bancos de dados foram apresentados e buscou-se ilustrar a utilização de tais bancos apontando possíveis ferramentas tecnológicas que poderão fazer uso de tais recursos. Espera-se que, ao avaliar a interface entre a Linguística e a Computação, este trabalho tenha contribuído com o debate teórico da Linguística e ao mesmo tempo ofereça instrumentos importantes para a implementação de recursos da linguagem utilizando o computador.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, L. *Um Estudo sobre Síntese de Fala para o Português Brasileiro*. 2005. Dissertação (Mestrado em Engenharia Elétrica). Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte.

ALVARENGA, D.; SOARES, M. B.; OLIVEIRA, M. A. de; NASCIMENTO, M. do. Da forma sonora da fala à forma gráfica da escrita uma análise linguística do processo de alfabetização. *Caderno de Estudos Linguísticos*, Campinas, n. 16, p. 5-30, jan./jun. 1989.

BOD, R.; HAY, J.; JANNEDY, S. (Eds.). *Probabilistic Linguistics*. Cambridge, Mass.: MIT Press, 2003.

BYBEE, J. *Morphology*: a study of the relation between meaning and form. Amsterdam, Philadelphia: John Benjamins, 1985.

_____. *Phonology and Language Use*. Cambridge: Cambridge University Press, 2001.

_____. *Language, Usage and Cognition*. Cambridge: Cambridge University Press, 2010.

_____.; HOPPER, P. (Eds.). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins, 2001.

CHOMSKY, N.; HALLE, M. *The Sound Pattern of English*. New York: Harper and Row, 1968.

CRISTÓFARO SILVA, T.; ALMEIDA, L. S. ASPA: a formulação de um banco de dados de referência da estrutura sonora do português contemporâneo. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, XXV 2005, São Leopoldo. *Anais*... São Leopoldo: Sociedade Brasileira de Computação, 2005. v. 1. p. 2268-2277. (CD-Rom).

_____.; GOMES, C. Representações múltiplas e organização do componente linguístico. *Fórum Linguístico* (UFSC), Florianópolis - Santa Catarina, v. 4, p. 147-177, 2007.

_____. ; OLIVEIRA, M. A. de . Variação do 'r' pós-consonantal no português brasileiro: Um caso de mudança fonotática ativada por cisão primária. *Letras de Hoje*, Porto Alegre, v. 37, p. 25-47, 2002.

_____.; MARTINS, R. M. F.; ALMEIDA, L. S.; OLIVEIRA-GUIMARAES, D. M. L. Alfabetização e conhecimento linguístico: o projeto e-Labore. In: SEVFALE, VI, 2007, Belo Horizonte. *Anais* ... Belo Horizonte: Faculdade de Letras - UFMG, 2007. v. 1. p. 1-16.

_____.; ALMEIDA, L. S.; MARTINS, R. M. F.; OLIVEIRA-GUIMARAES, D. M. L. Aquisição da escrita infantil: a construção de um corpus do português brasileiro. In: INTERNATIONAL JOINT CONFERENCE IBERAMIA/SBIA/SBRN, Workshop in Information and Human Language Technology, 4th, 2006, Ribeirão Preto. *Proceedings of the 4th Workshop in Information and Human Language Technology (TIL'2006). CD Room.* Ribeirão Preto: SBC, 2006.

GOLDSMITH, J. *Autosegmental and Metrical Phonology*. Oxford: Blackwell, 1990.

JOHNSON, K. Speech perception without speaker normalization, In: JOHNSON, K.; MULLENIX, J. (Eds.). *Talker variability in speech processing*. SanDiego: Academic Press, 1997. p. 146-165.

_____.; MULLENIX, J. (Eds.). *Talker variability in speech processing*. San Diego: Academic Press, 1997.

KAGER, R. *Optimality Theory*. Cambridge: Cambridge University Press, 1999.

KENSTOWICZ, M. *Phonology in Generative Grammar*. Oxford: Blackwell, 1994.

PIERREHUMBERT, J. Exemplar dynamics: Word frequency, lenition and contrast. In: BYBEE, J.; HOPPER, P. (Eds.). *Frequency and the emergency of linguistic structure*. Amsterdam: John Benjamins, 2001. p.137-157.

SAUSSURE, F. d. *Course in General Linguistics* (W. Baskin, Trans). New York: Philosophical Library, 1916.