ASPA: a Formulação de um Banco de Dados de Referência da Estrutura Sonora do Português Contemporâneo

Thaïs Cristófaro-Silva[1], Leonardo S. de Almeida[2], Thiago Fraga[2]

[1]Departamento de Letras – Universidade Federal de Minas Gerais (UFMG)
Laboratório de Fonetica (LABFON)

[2]Departamento de Engenharia Eletrônica – Universidade Federal de Minas Gerais (UFMG) – Centro de Estudos da Fala Acústica Linguagem e músicA (CEFALA)

thaiscristofarosilva@ufmg.br, almeida@cefala.org,
thfragasilva@yahoo.com.br

Abstract. *This article describes the formulation of the ASPA (Sound Assessment of Current Portuguese). This is an interdisciplinary project and was developed by researchers associated to LABFON and CEFALA laboratories. An automatic transcriber and parser for isolated words of Brazilian Portuguese where developed under the name LETRASON. Futhermore, the ASPA project offers a database which contains the categorization of the most frequent words in LAEL/PUC–SP Corpus. This database intends to be a great tool for studies in Probabilistic Phonology and Voice Synthesis.*

Resumo. *Este artigo descreve a formulação do projeto ASPA (Avaliação Sonora do Português Atual). O projeto ASPA é um projeto interdisciplinar desenvolvido pelo LABFON e pelo CEFALA. A equipe do projeto desenvolveu um transcritor sonoro e silabificador automático para palavras isoladas do português brasileiro, o LETRASON. Além disso, foi elaborado um banco de dados contendo a categorização das palavras de frequência mais significativa do corpus do LAEL/PUC–SP. Esse banco de dados pretende ser uma ferramenta de estudo disponibilizada para a comunidade científica interessada nas áreas de Fonologia Probabilística e Síntese de Voz.*

1. Introdução

O objetivo desse artigo é apresentar e discutir as bases da formulação do projeto ASPA: Avaliação Sonora do Português Atual.[1] A formulação inicial do Projeto ASPA deu-se a partir da necessidade de ter um conhecimento probabilístico sólido da estrutura sonora do português brasileiro contemporâneo. Tal necessidade decorreu basicamente de pressupostos teóricos de dois modelos que avaliam o componente sonora da fala: Fonologia de Uso [Bybee 2000, Bybee 2001] e Teoria de Exemplares [Johnson 1997, Pierrehumbert 2001, Pierrehumbert 2003]. Em linhas bastante gerais tais modelos postulam que a linguagem é compreendida como sendo multirepresentacional

[1]O projeto ASPA foi elaborado por Thaïs-Cristófaro-Silva(FALE-UFMG) em parceria com Leonardo Almeida (CPDEE-UFMG), Raquel Fontes-Martins (Poslin-FALE-UFMG) colabora como assistente de coordenação. O projeto conta ainda com a colaboração de César Reis (Labfon-FALE-UFMG), Hani Camille Yehia (Cefala-DELT-UFMG), Rafael Laboissiere (MaxPlank Institute-Germany) e Tony Sardinha (PUCSP).

sendo que o conhecimento é organizado em redes interligadas. O detalhe é crucial na organização do conhecimento. O conhecimento lingüístico é parte do conhecimento geral da espécie. A multirepresentacionalidade organizada implica num sistema individual, dinâmico e maleável (ao contrário dos modelos formais). Análises decorrentes desta proposta devem avaliar hipóteses sobre o conhecimento lingüístico e sobre o conhecimento em geral. Esta abordagem permite-nos explicar a diferenças individuais, as particularidades de grupos específicos e incorpora a dinamicidade presente em qualquer sistema/estrutura observada pela espécie humana.

O projeto ASPA é um empreendimento conjunto entre pesquisadores que atuam em áreas diversas do conhecimentos e que necessitam de um conhecimento sólido da organização sonora do português contemporâneo. O entrelace maior entre estes pesquisadores é a concepção teórica de que o conhecimento lingüístico é organizado probabilisticamente. Os resultados do Projeto ASPA oferecem subsídios a pesquisas em diversas áreas do conhecimento, dentre estas:

a. Teorias lingüísticas;
b. Teorias fonéticas e fonológicas;
c. Ensino de fonética e fonologia;
d. Lingüística de corpora;
e. Lingüística aplicada à educação;
f. Organização de banco de dados;
g. Lingüística computacional;
h. Formulação de software.

A originalidade do presente projeto: é sobretudo oferecer um instrumento de apoio a pesquisas que têm por objetivo avaliar os tipos fonológicos em corpus do português contemporâneo. O ASPA busca oferecer contribuições para a análise do mapeamento de tipos silábicos e segmentais do português brasileiro contemporâneo. Neste contexto tipos explicita qualquer categoria identificada no sistema sonoro.

A seguir, a seção 2 apresenta as linhas gerais da metodologia empregada. A seção 3 descreve as características básicas do projeto, as regras formuladas e a implementação tecnológica do mesmo. Finalmente, a conclusão indica os próximos passos a serem seguidos pela equipe do projeto a fim de oferecer a comunidade acadêmica um subsídio sólido para os estudos relacionados a sonoridade do português contemporâneo.

2. Metodologia

2.1. Corpus de Análise

O Projeto ASPA adotou como corpus de análise os dados do Projeto DIRECT–PUC-SP: `http://lael.pucsp.br/direct`. Mais especificamente consideramos dados do corpus de escrita que nos foi encaminhado em 2004.[2] Nos referiremos a tal material como Corpus do LAEL. Uma avaliação em detalhes da formulação e desenvolvimento do corpus do LAEL pode ser obtida em [Sardinha 2003]. O corpus é composto por um total de um total de 607.392 palavras (ou tipos) que totalizam 228.766.402 tokens.

[2]Registramos aqui o nosso agradecimento à equipe do Project Direct e em especial ao Professor Tony Sardinha que intermediou o nosso contato.

Optamos pelo corpus de escrita por duas razões principais. A primeira delas é que há maior proximidade com a forma ortográfica prescrita e sendo assim a transcrição sonora se torna mais eficiente. A segunda razão para optarmos pelo corpus de escrita é a sua dimensão que é significativamente maior do que o corpus de fala oferecendo maior diversidade de tipos, ou seja, palavras a serem consideradas.

Visando a operacionalidade do trabalho de transcrição optamos por transcrever inicialmente um total de 199.864 palavras (do total de 607.392 do corpus integral). Desta maneira foram cadastradas no ASPA desde a palavra de maior frequência de ocorrência no português "de" com 10.739.395 ocorrências, até a palavra "Zylium" que é a última a apresentar frequência de ocorrência 6 no corpus do LAEL. Sendo assim, palavras que possuem freqüência de ocorrência menor que 5 não foram incluídas nesta etapa inicial de cadastro e transcrição sonora dos dados do ASPA.

2.2. Categorias Cadastradas

Cada palavra listada no corpus do LAEL que foi utilizado no Projeto ASPA teve as seguintes informações cadastradas pelos colaboradores:

1. Ortografia.
2. Transcrição sonora.
3. Freqüência de ocorrência.
4. Categoria gramatical: adjetivo advérbio, artigo, conjunção, interjeição, preposição, pronome, numeral, substantivo, verbo, dúvida. Quando pertinente mais de uma categoria gramatical foi cadastrada. Por exemplo, "amada" foi cadastrada como adjetivo, verbo e substantivo.
5. Morfologia: flexionado (verbal), flexionado (plural), derivado, original (sem flexão ou derivação), dúvida. Quando pertinente mais de uma categoria para morfologia foi cadastrada. Por exemplo, "atividades" foi cadastrada como flexionado (plural) e derivado.
6. Origem: africana, indígena, dúvida. Tal categoria foi inserida para oferecer a oportunidade de avaliação do contato lexical das línguas nativas e africanas no português do Brasil. O contato lingüístico entre povos nativos e povos africanos com os portugueses é compreendido por alguns autores como um processo de crioulização do português brasileiro (o que não ocorreu em larga escala com o português europeu).
7. Data da inclusão e colaborador responsável que efetuou o cadastro (dado não disponibilizado ao público).

2.3. Transcrição Sonora

Na figura 1 pode-se observar que os dados ortográficos foram transcritos com símbolos fonéticos. Esta transcrição foi realizada a partir do software desenvolvido por Leonardo Almeida em parceria com Thaïs Cristófaro-Silva que converte símbolos ortográficos do português em seus correlatos sonoros que são representados graficamente por símbolos do Alfabeto Internacional de Fonética [3] e alguns símbolos gráficos que serão apresentados a seguir. Denominamos o aplicativo responsável pela transcrição ortográfica-sonora de LETRASON.

[3] `http://www2.arts.gla.ac.uk/IPA/ipa.html`

Não temos a pretensão de avaliar a interface fonética-fonologia superficialmente neste documento. Tal discussão fica sugerida para eventos futuros. Contudo, é importante ressaltar que a transcrição gerada pelo LETRASON oferece maior informação de detalhe fonético do que é tradicionalmente assumido nas análises fonológicas tradicionais [Cristófaro-Silva 2001, Mateus 1975, Mattoso-Câmara 1975]. Ao mesmo tempo a transcrição não inclui todo e qualquer detalhe fonético atestado no português brasileiro. O nosso trabalho foi de, mesmo diante das adversidades que a tarefa impõe, sugerir uma transcrição sonora que permita a busca de elementos sonoros em qualquer variedade lingüística do português brasileiro. Neste sentido sugerimos a denominação "transcrição sonora" para os dados gerados pelo LETRASON que são graficamente representados por símbolos do Alfabeto Internacional de Fonética e alguns outros símbolos adicionais.

2.3.1. Símbolos Vocálicos

Um conjunto de 15 símbolos vocálicos são utilizados no ASPA. Incluem-se dentre estes as sete vogais orais que ocorrem em sílaba tônica /i,ɛ,e,a,ɔ,u/ ; cinco vogais nasais que são obrigatoriamente nasais no português /ĩ,ẽ,ã,õ,ũ/ e adicionalmente os símbolos /E/ e /O/ são utilizados para representar as vogais médias pretônicas que podem se manifestar como abertas, fechadas ou alçadas ("perigo, bonito"). Além disso utiliza-se uma vogal epentética /I/ que ocorre entre determinados encontros consonantais ("advogado, pneu"). A Tabela 1 ilustra cada um destes símbolos vocálicos indicando um exemplo do português.

Tabela 1. Exemplos de palavras com os símbolos vocálicos utilizados.

Símbolo	Exemplo	Símbolo	Exemplo
/i/	v[i]da	/ĩ/	s[ĩ]to
/e/	[e]xito	/ẽ/	s[ẽ]pre
/ɛ/	bon[ɛ]	/ã/	s[ã]ba
/a/	c[a]sa	/õ/	t[õ]to
/o/	av[o]	/ũ/	n[ũ]ca
/ɔ/	c[ɔ]po	/E/	p[E]rigo
/u/	sa[u]de	/O/	pr[O]cura
/I/	ad[I]vogado		

2.3.2. Ditongos

Os ditongos foram representados por um dos símbolos vocálicos listados acima e um símbolo que indica a vogal assilábica do ditongo. A vogal assilábica alta anterior foi transcrita como /j/ e a vogal assilábica alta posterior foi transcrita por /w/. Alguns exemplos de transcrição de ditongos orais e nasais são apresentadas na Tabela 2.

Tabela 2. Exemplos de palavras com os símbolos de ditongo utilizados.

Símbolo	Exemplo
/j/	cu[j]dado
/w/	ca[w]da

2.3.3. Consoantes

Nos casos em que ocorre uma oclusiva velar seguida de vogal assilábica posterior utilizamos o símbolo de uma consoante complexa /kʷ,gʷ/. Além das duas consoantes complexas listadas anteriormente o ASPA utiliza 25 consoantes. Dentre essas 25 consoantes existem 19 que representam as consoantes que são tradicionalmente utilizadas no português brasileiro /b,k,d,f,g,ʒ,l,ʎ,m,n,ɲ,p,ɾ,h,s,ʃ,t,v,z/ e 6 outros símbolos /tʃ,dʒ,R,L,S,Z/. Exemplos com os símbolos utilizados são apresentados na Tabela 3.

Tabela 3. Exemplos de palavras com os símbolos de consoantes utilizados.

Símbolo	Exemplo	Símbolo	Exemplo	Símbolo	Exemplo
/kʷ/	[kʷ]atro	/ʎ/	pa[ʎ]a	/t/	ca[t]arro
/gʷ/	á[gʷ]a	/m/	ca[m]ada	/v/	[v]alor
/b/	[b]ola	/n/	ca[n]eta	/z/	[z]ebra
/k/	[k]rime	/ɲ/	fari[ɲ]a	/tʃ/	[tʃ]eco
/d/	[d]ado	/p/	[p]ato	/dʒ/	lin[dʒ]a
/f/	[f]ivela	/ɾ/	pa[ɾ]ada	/R/	ca[R]ta
/g/	[g]arfo	/h/	a[h]oz	/L/	ca[L]do
/ʒ/	a[ʒ]uda	/s/	[s]ela	/S/	fe[S]ta
/l/	[l]ata	/ʃ/	[ʃ]á	/Z/	a[Z]ma

2.4. Silabificação

Todos as palavras após serem transcritas foram silabificadas automaticamente e verificadas pelos colaboradores. A cada uma das sílabas foi atribuido um valor de tonicidade. Caracterizou-se a sílaba tônica, postônica medial, postônica final, e pretônicas desde a mais próxima da sílaba tônica até a mais distante da sílaba tônica. Como a caracterização de tonicidade foi numérica será possivel selecionar cabeças de pés métricos pretônicos bem como pés degenerados.

2.5. Dados Excluídos

Alguns dados encontrados no corpus do LAEL não se adequaram ao mapeamento sonoro que sugerimos. Tais dados foram agrupados em categorias específicas e poderão ser consultados por usuários do ASPA sendo que seu número de listagem e a sua freqüência de ocorrência é preservada. Os dados foram agrupados como na Tabela 4.

2.6. Problemas Metodológicos

Obviamente que nos deparamos com inúmeras adversidades no desenvolver do projeto. Alguns dos problemas metodológicos são apresentados a seguir.

Tabela 4. Categorias de exclusão de palavras.

Classificação	**Exemplo**	Freqüência
Siglas	FHC	105,830
Pronúncia não inferível Escrita não inferível (pelo conversor) em relação a ortografia do português, com pronúncia instável e geralmente estrangeira.	ZERBETTO	6
Pronúncia instável Empréstimo com pronúncia instável.	BUNCHEN	2
Erro gráfico A ortografia está incorreta e a inferência pode ser dúbia/problema.	MAMOBRA	1
Língua inglesa Palavras do Inglês.	MACCLELLAND	2
Outras línguas Palavras de línguas diferentes do inglês.	TOUR	3,864

1. Duas grafias para uma mesma palavra. A grande maioria dos casos diz respeito a ausência/presença de acento gráfico: saída e saida. Nestes casos as duas formas ortográficas diferentes recebem a mesma transcrição e serão listadas no arquivo de buscas solicitada pelo usuário do ASPA.
2. Problemas relacionados a impossibilidade de identificação da pronúncia. Para estes casos sugere-se que o pesquisador faça uma busca geral nos dados totais do LAEL:
   a. Alternância vocálica em nomes e verbos: o esb[o]ço, eu esb[ɔ]ço;
   b. Acentuação: oscar-oscár , recorde-recórde;
   c. Nomes cuja grafia potencializa duas pronúncias: s[e]de-s[ɛ]de; f[o]rma-f[ɔ]rma;
   d. Casos potenciais de ditongo-hiato (optou-se por preservar o hiato): maizena, saideira, juizado;
   e. Casos de potencial epêntese foram analisados como uma vogal epentética: dogma, afta, técnica.

3. Implementaçao Tecnológica

A construção de um banco de dados com cerca de 200.000 palavras transcritas foneticamente e divididas em sílabas requer o desenvolvimento de um software de transcrição automática [Gomes 1998]. Caso contrário, a equipe responsável pela construção de tal banco de dados perderia meses ou até mesmo anos realizando a transcrição e a divisão em sílabas de cada palavra. Portanto, um software de transcrição automática, o LETRASON, foi desenvolvido pela equipe do projeto ASPA. Além disso, também foi desenvolvido um algoritmo simples de silabifacao automática capaz de processar a transcrição obtida com o LETRASON. Os dois aplicativos foram desenvolvidos utizando a linguagem C.

Após as etapas de transcrição e silabificação automática, foi desenvolvida uma página na Internet, utilizando-se a linguagem PHP. Essa página foi acessada pelos colaboradores do projeto ASPA que eram responsáveis por adicionar informações de categoria gramatical, morfologia, origem e tonicidade a cada uma das palavras.

3.1. LETRASON

O desenvolvimento de um software capaz de transcrever palavras isoladas não é muito trivial. Além das dificuldades já discutidas na seção 2.3. alguns problemas ocorrem na realização desta tarefa [Dutoit 2001]:

a. Um único caracter pode corresponder a mais de um fonema, como na palavra "aptidão" onde o caractere p deve ser transcrito como [pɪ];
b. Uma sequência de caracteres pode corresponder a um único fonema. Na palavra "chá", por exemplo, os dois primeiros caracteres são transcritos como apenas um fonema [ʃ].
c. Um caractere pode não corresponder a nenhum fonema, como o h da palavra "hoje".
d. O mesmo caractere pode ser ser transcrito de duas maneiras diferentes dependendo dos caracteres que o precedem e o que seguem. Por exemplo, o s nas palavras "casca" e "asma" deve ser transcrito como [s,ʃ] e [z,ʒ] respectivamente.

Sabendo-se de todas estas dificuldades iniciais, a primeira atitude a ser tomada no desenvolvimento de um transcritor automático é a elaboração de um código que seja capaz de representar todos os fonemas do português. A opção imediata seria utilizar as fontes IPA (International Phonetic Alphabet), porém, apesar de tais fontes serem compatíveis com a maioria dos processadores de texto atuais elas não são compreendidas pelas principais ferramentas de programação utilizadas no presente projeto. Portanto, decidiu-se criar um código de quatro letras (Tabela 5) para cada fonema do português brasileiro. Cada letra deste código representa uma informação sobre o fonema. A primeira letra classifica o fonema como consoante ou vogal. A segunda, terceira e quarta letra possuem diferentes significados para consoantes e vogais. Para consoantes, a segunda letra diz respeito ao modo de articulação, a terceira ao local de articulação e a quarta ao vozeamento ou não do fonema. Para as vogais, a segunda e a terceira letra fornecem informação a respeito da altura da língua e quarta letra diz respeito a posição da língua em relação ao trato vocal.

Tabela 5. Exemplos de códigos utilizados na transcrição automática.

Fonema	Código	Descrição
[p]	COBD	Consoante Oclusiva Bilabial Desvozeada
[t]	COAB	Consoante Oclusiva Alveolar Desvozeada
[ɲ]	CNPV	Consoante Nasal Palatal Vozeada
[ɔ]	VMBP	Vogal Média-Baixa Posterior

Logo após definirem-se os códigos para cada fonema do português brasileiro foi necessário criar um conjunto de regras que fazem o mapeamento de caracteres de palavras isoladas em fonemas. Essas regras se utilizam de informações contidas nas palavras isoladas, sendo assim, elas conseguem realizar a transcrição fonética levando em conta apenas as ordens dos caracteres. O programa de transcrição automática processa cada caractere da palavra e aplica a regra específica para tal caractere. Por exemplo, ao encontrar um caractere t em uma determinada palavra o programa executa a regra de transcrição específica: se o caracter t não for seguido das consoantes(b,ç,d,f,g,j,m,n,p,s,t,x,z) ele deve ser transcrito como [t] caso contrário ele deverá ser transcrito como [tɪ]. A parte do código do LETRASON que implementa essa regra é descrita abaixo:

```
/* Regra de transcrição do caractere 'c' */
if ( (palavra[i+1] == 'b') || (palavra[i+1] == 'ç')
   ||(palavra[i+1] == 'd') || (palavra[i+1] == 'f')
   ||(palavra[i+1] == 'g') || (palavra[i+1] == 'j')
   ||(palavra[i+1] == 'm') || (palavra[i+1] == 'n')
   ||(palavra[i+1] == 'p') || (palavra[i+1] == 's')
   ||(palavra[i+1] == 't') || (palavra[i+1] == 'x')
   ||(palavra[i+1] == 'z') )
{/* t seguido (b,ç,d,f,g,j,m,n,p,s,t,x,z)*/
 /*Atualiza o vetor que armazena os códigos de 4 palavras*/
  codigo[k] = 'C';
  codigo[k+1] = 'O';
  codigo[k+2] = 'A';
  codigo[k+3] = 'D';
  codigo[k+4] = '\0';
  codigo[k+5] = 'V';
  codigo[k+6] = 'A';
  codigo[k+7] = 'E';
  codigo[k+8] = 'A';
  codigo[k+9] = '\0';
  k += 10;/* Adianta a posição da memória no vetor código */
  i++;/* Faz o programa transcrever o caracter seguinte */
}
else{  /* regra t */
  /*Atualiza o vetor que armazena os códigos de 4 palavras*/
   codigo[k] = 'C';
   codigo[k+1] = 'O';
   codigo[k+2] = 'A';
   codigo[k+3] = 'D';
   codigo[k+4] = '\0';
   k += 5;/* Adianta a posição da memória no vetor código */
   i += 1;/* Faz o programa transcrever o caracter seguinte */
}
```

Sendo assim, o programa de transcrição automática simplesmente realiza uma série de decisões se-então para cada caractere encontrado na palavra. Para realizar a transcrição de todos os caracteres em fonemas foi elaborado um conjunto de noventa e duas regras. Estimativas em testes preliminares indicam que esse conjunto de regras é capaz de transcrever com precisão mais de 90% das palavras em português presentes no corpus do LAEL.

3.2. Silabificador Automático

Após a transcrição da palavra em uma seqüência de códigos de quatro letras que representam fonemas é possível dividí-la em sílabas utilizando o software de silabificação automática. Este software lê a sequências de fonemas transcritos e consegue dividir a palavra em sílabas utilizando-se de apenas 6 regras. Estas regras se baseiam apenas na informação contida nos códigos de quatro letras e nas ordenações dos mesmos. Estimati-

vas em testes premilinares indicam que as regras de silabificação são eficazes em mais de 99% das palavras transcritas utilizando o software de trascrição descrito acima.

3.3. Cadastro de Palavras

A equipe do projeto ASPA cadastrou uma a uma as 199.864 palavras mais freqüentes do corpus do LAEL. Ao cadastrar uma palavra cada coloraborador era responsável por adicionar informações de categoria gramatical, morfologia, origem e tonicidade. Para facilitar o trabalho dos colaboradoes foi desenvolvida a página do cadastro do projeto ASPA(exemplo: `www.projetoaspa.org/cadastro/teste.html`). Para acessar a página o colaborador entrava com um login e uma senha que lhe possibilitavam acesso ao seu lote de palavras. Cada lote de palavras era composto por 6000 palavras que eram mostradas na tela do navegador em blocos de 10. Ao terminar o cadastro de um bloco o colaborador apertava um botão responsável por enviar o cadastro ao servidor do projeto. Em caso de dúvida de categorização, discordância com a transcrição ou palavras que se encaixam na descrição da seçao 2.5. o colaborador poderia marcar um botão do tipo “tick mark” para a palavra que seria então enviada para análise futura. É importante ressaltar que todos os colaboradores passaram por uma etapa de treinamento, onde eram acompanhados pelos coordenadores do projeto. Esta etapa de treinamento foi importante por garantir a maior uniformidade possível ao cadastro de palavras.

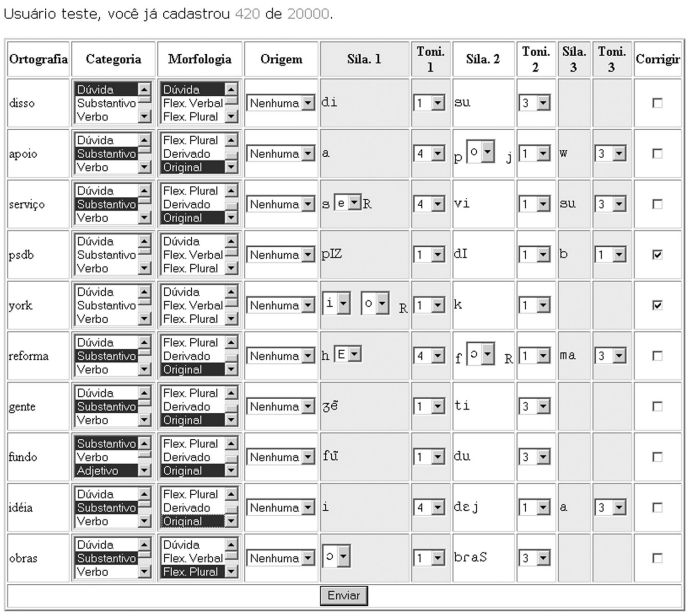

Figura 1. Exemplo de tela de cadastro

4. Conclusão

Neste artigo foram descritos os objetivos e a metodologia do projeto ASPA. O transcritor automático descrito, LETRASON, apresenta um excelente desempenho na transcrição do português brasileiro. Porém, para ser utilizado em síntese de fala ele deve sofrer algumas modificações que permitam a incorporação de regras que o tornem capaz de transcrever corretamente palavras conectadas. Por exemplo, atualmente a expressão “bolas amarelas” é transcrita como /bɔ-laS a-ma-rɛ-las/ quando na realidade deveria ser transcrita como /bɔ-la-za-ma-rɛ-las/).

No entanto, a transcrição realizada pelo LETRASON (códigos de 4 letras) possui algumas vantagens. A primeira é que um conjunto de apenas 6 regras de silabificação consegue particionar quase todas as palavras transcritas, essa característica é muito interessante para sistemas que utilizem síntese de voz por concatenação de sílabas. Além disso, o código de 4 letras utilizado é bastante útil na formulação de um banco de dados que possibilite busca segmental em suas palavras. Deste modo, por exemplo, pode se realizar buscas no banco de dados do ASPA que sejam capazes de responder a perguntas do tipo: Em quantas e em quais palavras do português brasileiro existe consoante fricativa em final de sílaba seguida por consoante oclusiva?

O banco de dados do projeto ASPA é ainda mais completo. Pois, além de informação sonora, as palavras cadastradas ainda possuem informação de tonicidade, morfologia, categoria, origem e freqüência, e sendo assim, as buscas realizadas em sua base de dados podem ser bastante refinadas. É importante ressaltar que a informação a respeito da tonicidade de cada sílaba pode ser convenientemente utilizada na elaboração de modelos prosódicos para sistemas de síntese de fala. Portanto, quando todo o banco de dados do ASPA estiver disponível para busca eletrônica na Internet ele contribuirá para diversas áreas da ciência da fala.

Referências

Bybee, J. (2000). The phonology of the lexicon: evidence from lexical diffusion. In Barlow, M. and Kemmer, S., editors, *Usage-based models of language*, pages 65–85. CSLI Publications.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press.

Cristófaro-Silva, T. (2001). *Fonética e Fonologia do Português*. Editora Contexto.

Dutoit, T. (2001). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers.

Gomes, L. C. T. (1998). *Sistema de Conversão Texto-Fala para a Língua Portuguesa utilizando a abordagem de Síntese por Regras*. Tese de Mestrado - FEEC/UNICAMP.

Johnson, K. (1997). Speech perception without speaker normalization: an exemplar model. In Johnson, K. and Mullenix, J. W., editors, *Talker variability in speech processing*, pages 145–165. San Diego: Academic Press.

Mateus, M. H. M. (1975). *Aspectos da Fonologia Portuguesa*. Centro de Estudos Filológicos, 19.

Mattoso-Câmara, J. (1975). *História e Estrutura da Língua Portuguesa*. Editora Padrão, 2a Edição.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. and Hopper, P., editors, *Frequency effects and the emergence of linguistic structure*, pages 1–19. John Benjamins.

Pierrehumbert, J. (2003). Probabilistic phonology: discrimination and robustness. In R. Bod, J. Hay, S. J., editor, *Probabilistic linguistics*, pages 177–228. MIT Press.

Sardinha, T. B. (2003). The bank of portuguese. *Direct Papers*, 50.