Transcrição fonética automática para lemas em verbetes de dicionários do Português do Brasil

Vanessa Marquiafável, Claudia Zavaglia

Departamento de Letras Modernas do Instituto de Biociências, Letras e Ciências Exatas de São José do Rio Preto (IBILCE-UNESP) Rua Cristóvão Colombo, 2265, Jardim Nazareth, CEP. 15054-000, São José do Rio Preto - SP - Brasil

marquiafavel@gmail.com, zavaglia@ibilce.unesp.br

Abstract. *This paper describes the steps for building a computational environment to support automatic phonetic transcription of lemmas for Brazilian dictionaries entries. Our main assumption is that from a set of computer applications it would may be possible to convert characters of isolated lexical units (lemmas) in their corresponding phonetic units, using a set of objective criteria for annotation without any kind of human intervention.*

Resumo. *Este artigo descreve as etapas para a construção de um ambiente computacional para suporte à transcrição fonética automática para lemas de verbetes em dicionários do Português do Brasil. Nossa hipótese norteadora é a de que a partir de um conjunto de aplicações computacionais poderá ser possível converter caracteres de unidades lexicais isoladas (verbetes) em suas unidades fonéticas correspondentes, com a aplicação de um conjunto fixo e objetivo de critérios de anotação, sem que haja para isso qualquer tipo de intervenção humana.*

1. Introdução

A anotação linguística em *corpora* pode se dar nos níveis fonético, morfológico, sintático, semântico, pragmático e discursivo, e pode ser realizada manualmente (por linguistas), automaticamente (por ferramentas de PLN) ou semiautomaticamente (pósedição manual da saída de uma ferramenta de PLN). A lista de etiquetas a ser utilizada em cada um desses níveis dependerá: das características linguísticas do *corpus*, dos objetivos da anotação, dos limites do programa e dos pressupostos teóricos aplicados (Garside, Leech e McEnery 1997).

Tanto na área de Linguística quanto nas áreas de Tecnologia da Fala, as transcrições fonéticas são frequentemente necessárias, embora seja uma tarefa demorada, repetitiva e custosa, contribuindo para que erros humanos possam ocorrer natural e frequentemente (Cucchiarini, 1993; Shriberg & Lof, 1991). No entanto, a automatização desse tipo de procedimento torna possível a anotação fonética de pequenos ou grandes *corpora* por meio da aplicação de um conjunto fixo e objetivo de critérios de anotação. Sem que haja intervenção humana, é produzida uma transcrição sem subjetividades. Garante que diferentes resultados produzidos em diferentes momentos sejam idênticos, quando produzidos com a mesma ferramenta computacional. Além disso, as transcrições automáticas podem ser criadas com uma fração do custo e do tempo com que são geradas as transcrições manuais. Frente a isso, diversos estudos têm se dedicado à

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 154–158, Cuiabá, MT, Brazil, October 24–26, 2011.

criação de sistemas computacionais que automatizem a tarefa de transcrição fonética, embora em número bem menor dedicados ao português do Brasil (doravante PB), quando comparados ao Inglês (Siravenha *et al*, 2008). Vale ainda dizer, que a grande maioria dos grupos de pesquisa realiza a construção de transcritores fonéticos automáticos para uso local e particular de seus membros, em geral, para sistemas de síntese e reconhecimento de fala.

Nesse contexto, este projeto de pesquisa surge com o intuito de construir um ambiente Web gratuito de suporte à transcrição fonética automática para lemas em verbetes de dicionários do PB. Nossa hipótese norteadora é a de que a partir de um conjunto de aplicações computacionais será possível converter caracteres de unidades lexicais isoladas (os lemas da nomenclatura de um dicionário) em suas unidades fonéticas correspondentes. Se confirmada, os lexicógrafos, nosso principal público-alvo, poderão desfrutar dos benefícios da automatização da transcrição fonética de verbetes em dicionários, enriquecendo-os e diminuindo o tempo e as dificuldades de inserção desse tipo de informação em suas obras lexicográficas. Além disso, sua disponibilização gratuita via Web permitirá que demais interessados em transcrição fonética automática de palavras isoladas possam também aproveitar as vantagens que esse tipo de sistema poderá oferecer.

Além da Introdução, este trabalho apresenta outras duas partes. Na primeira é apresentada a metodologia proposta para o desenvolvimento do transcritor fonético almejado. Há também a apresentação do seu funcionamento e da construção de um *corpus* que auxiliará tanto para seu treinamento quanto para sua avaliação. Ainda nessa mesma parte, é citada a tecnologia existente atualmente para a implementação computacional da ferramenta Na segunda parte temos nossa conclusão, com base nas expectativas que pretendemos alcançar, dado que esta pesquisa iniciou-se em março de 2011.

2. Metodologia Proposta

2.1 Esquema de funcionamento

O processo de transcrição fonética será realizado da seguinte maneira: para cada segmento (unidade ortográfica/grafema) do lema, o conjunto de regras fonéticas será consultado, partindo da esquerda para a direita, até que uma regra apropriada seja encontrada e aplicada. A seguir, o próximo segmento será analisado e o processo se repetirá até que toda a palavra esteja transcrita. As regras fonéticas a serem utilizadas para a geração automática do dicionário fonético serão elaboradas a partir da transcrição manual de um *corpus*. Nele estarão contidos o maior número possível de sons do PB, e sua transcrição será supervisionada por um especialista da área de Fonética e Fonologia.

2.2 *Corpus* para construção das regras fonéticas e avaliação da ferramenta

Para a criação da lista de regras fonéticas, com vistas à sua implementação computacional, e considerando a necessidade de nosso público-alvo, adotaremos o sistema IPA (*International Phonetic Alphabet*), uma vez que apresenta completa representação fonética dos sons das línguas naturais. A fonte escolhida para a transcrição fonética dos dados é a *Arial Unicode MS*, que é a mais adequada a esse tipo de aplicação por se servir de caracteres *Unicode*, legíveis por computador. O ideal seria a construção de um conjunto de regras linguísticas com todas as pronúncias aceitáveis do PB e, de

preferência, indicando o estilo e as propriedades de cada uma. Tal abordagem, obviamente, não é plausível, devido à dimensão de um inventário dessa proporção, o que nos traria ainda o problema de definir qual o modelo de pronúncia que deveria estar contido no *corpus*. Embora a tendência predominante na atualidade (Cremelie e Martens, 1999; Quilis (1982)) seja a de refletir um modelo de dicção o mais universal possível, sem variedades linguísticas, podemos observar que isto é impossível. Isso porque as variedades do PB se diversificam conforme as regiões geográficas, as classes sociais, os níveis de escolarização, as idades e assim por diante. Um exemplo de trabalho que objetiva fazer um levantamento das variedades do PB é o projeto do Atlas Linguístico do Brasil[1]. Dessa maneira, entre todas as variedades linguísticas existentes, adotaremos o dialeto paulista. Por dialeto paulista deve-se entender a fala de pessoas cultas oriundas do Estado de São Paulo. Essa escolha foi motivada pelo fato de tanto os orientadores quanto a orientanda serem nativos dessa variedade e, portanto, facilitar o processo de construção do conjunto das regras fonéticas existentes na variedade adotada.

Para podermos avaliar a eficiência da ferramenta de transcrição, será necessária a construção de um *corpus* foneticamente anotado, que servirá como nosso *corpus goldstandard*. Ele será composto por uma lista de palavras e suas respectivas pronúncias. Já que o conjunto de palavras de uma língua é um conjunto aberto (devido aos neologismos e composições que surgem constantemente) e, dado os recursos limitados deste projeto (tempo e tamanho de equipe), um critério para a seleção de palavras para esse *corpus* teve que ser adotado, no caso, o frequencial. Para isso, poderão ser utilizadas as palavras mais frequentes contidas nos *corpora* do PB disponíveis gratuitamente pelo Projeto AC/DC[2]. Isto porque são extraídas de variados gêneros textuais (científico, jornalístico, etc.) e passaram por uma vistoria rigorosa quanto à sua qualidade ortográfica, por exemplo. Sardinha (2004) reforça essa vantagem de reaproveitamento dos recursos de grandes *corpora*, além de o usuário não precisar ter todo o trabalho de coletar um *corpus* novo; essa reutilização contribui para a condição de existência de dados verificáveis, não comprometendo a pesquisa em termos de replicabilidade e generalidade. Depois de delimitado o *corpus*, o próximo passo será a transcrição fonética manual de cada uma das palavras dessa lista. Para isso, utilizaremos o editor de texto *X-Emacs*[3], pois possibilita a manipulação de grande quantidade de dados sem que haja travamento de seu sistema.

2.3 Sistemas baseados no conhecimento

Entre as técnicas de aprendizado de máquina encontradas na literatura, temos: (i) a baseada em regras, geralmente amparada por um linguista, que faz uso de regras prédefinidas para a conversão da forma ortográfica e do respectivo símbolo sonoro; (ii) o algoritmo *data-driven*, que gera automaticamente as regras de transcrição a partir de um dicionário de treino, com o auxílio de expressões regulares; e (iii) a mista, em que as regras linguísticas e/ou estatísticas são geradas com base em transcrições fonéticas.

Neste trabalho, pretendemos mostrar como os pressupostos teóricos da Fonética

1 Mais detalhes em http://twiki.ufba.br/twiki/bin/view/Alib/WebHome.

2 O projeto AC/DC (Acesso a corpos/Disponibilização de corpos), iniciado em 1999, surgiu da necessidade de juntar os poucos recursos disponíveis num único ponto na rede e dessa forma facilitar a comparação e a reutilização do material, permitindo ao mesmo tempo acesso a uma ferramenta poderosa de interrogação de corpos, o sistema CWB, para o qual a Linguateca desenvolveu esta interface. Mais informações vide http://www.linguateca.pt/ACDC .

3 Mais informações e download gratuito do programa acessar http://www.xemacs.org/

podem contribuir para o sucesso de nosso sistema de conversão ortografia-fone(ma), utilizando para isso uma linha de programação de algoritmos segundo regras linguísticas do PB. A Linguística, quando aplicada ao processamento computacional transcritivo, permite construir sistemas, em princípio, mais leves (não necessitam de grandes bases de dados) e eficientes, (pois consomem menos memória durante o processamento). Serve-se de listas de regras altamente dependentes de linguista, cuja presença será obrigatória tanto no desenvolvimento quanto na supervisão dos dados. Os sistemas baseados em regras são mais fáceis de implementar computacionalmente e seu formato permite fazer simples modificações. Essas regras refletem o conhecimento linguístico implicado na conversão além de permitir avaliar a adequação das mesmas em grandes bases de dados, estudar os erros por elas produzidos e corrigi-los. Apesar de ter a desvantagem de quanto maior o número de regras, menor a velocidade de execução, este é ainda um problema que pode ser resolvido com os avanços informáticos existentes. A abordagem de regras linguísticas construída a partir do conhecimento fonético-fonológico tem sido recentemente aplicada a sistemas de conversão texto-fala em Português Europeu (Braga et al. 2006) e Português do Brasil (Silva et al. 2010), com êxito comprovado.

Após sua elaboração, serão feitas avaliações sobre a qualidade do material transcrito automaticamente. Tais avaliações também servirão para ajustes na precisão de anotação da ferramenta proposta. Para avaliar o desempenho de nosso transcritor, compararemos as múltiplas transcrições por ele efetuadas com aquelas elaboradas por 2 linguistas especialistas (orientanda e coorientador) para a construção das regras de transcrição. Consideraremos que o nosso sistema gerará automaticamente transcrições fonéticas de qualidade quando o percentual de concordância entre as transcrições dos especialistas e as geradas automaticamente estiver muito próximo de 100%.

3. Conclusão

Com a disponibilização gratuita de um transcritor fonético automático poderemos: 1) otimizar a tarefa minuciosa, morosa e cansativa de transcrição fonética manual de *corpora*; 2) possibilitar que a ferramenta proposta seja utilizada também como um segundo anotador nas situações de pesquisa em que a transcrição fonética é feita por apenas uma pessoa; 3) permitir que seja utilizada em situações de anotação fonética de *corpora* nas quais mais de um anotador está envolvido, com o intuito de se solucionar possíveis dúvidas sobre o que realmente deveria ter sido transcrito, segundo o protocolo de anotação adotado. Tal artifício é comprovadamente muito mais barato do que encontrar outra pessoa para desempenhar esse papel de 'juiz'; 4) eliminar parte da subjetividade em transcrições de grandes *corpora*, originada pela impossibilidade de se obter uniformidade na obediência aos critérios de anotação fonética adotados durante a transcrição de *corpora* de grandes dimensões, uma vez que quanto mais pessoas são necessárias para transcrever um mesmo *corpus*, maior pode ser a falta de uniformidade nos critérios aplicados; 5) dedicar nossos esforços também à diminuição da deficiência em estudos sobre transcrição fonética automática para o PB; 6) e para a Lexicografia, nosso principal objetivo, oferecer uma ferramenta de anotação fonética automática, que poderá gerar facilmente e em poucos instantes listas fonéticas a partir de uma entrada em formato ortográfico. Isso, sem as dificuldades mencionadas na Introdução deste, facilitando assim a inserção de informação fonética em todas as obras que produz, enriquecendo-as e diminuindo o tempo e as dificuldades de inserção desse tipo de informação.

4. Referências

Braga, D.; L. Coelho & F. G. V. Resende Jr. (2006). “A rule-based grapheme-to- Phone converter for TTS systems in European Portuguese”. In VI International Telecommunications Symposium (ITS2006), Setembro de 2006. Fortaleza-CE, Brasil, 976-981.

Cremelie, N.; Martens, J.P. (1999) In “Search of better pronunciation models for speech recognition”. Speech Communication, 29, pp. 115-136.

Cucchiarini, C. (1993) “Phonetic transcription: a methodological and empirical study”. PhD Thesis, University of Nijmegen.

Garside, R., Leech, G. N., and McEnery, T. (1997). Corpus annotation: linguistic information from computer text corpora. London: Longman.

Quilis, A. (1982) “Diccionarios de pronunciación”, Lingüística Española Actual IV,II: 326-332.

Sardinha, T. B. (2004). “Linguística de Corpus”. Barueri, SP: Manole.

Shriberg, L.D. And Lof, L. (1991) “Reliability studies in broad and narrow phonetic transcription”, Clinical Linguistics and Phonetics, 5, 225-279.

Silva, P.; Batista,P.; Neto,N.; Klautau,A. (2010). "An Open-Source Speech Recognizer for Brazilian Portuguese with a Windows Programming Interface". In: The International Conference on Computational Processing of Portuguese (PROPOR), Porto Alegre.

Siravenha, A.C.; Neto, N; Macedo, V; Klautau, A. (2008) “Uso de Regras Fonológicas com Determinação de Vogal Tônica para Conversão Grafema-Fone em Português Brasileiro”. In 7th International Information and Telecommunication Technologies Symposium.