Descrição do português com auxílio de programa computacional de interface

Leonor Scliar-Cabral[1], Vera Vasilévski[2]

[1] Laboratório de Produtividade Linguística Emergente (LAPLE) – Universidade Federal de Santa Catarina/CNPq – Florianópolis – SC – Brasil

[2] LAPLE – Universidade Federal de Santa Catarina/CAPES – Florianópolis – SC – Brasil

lsc@th.com.br, vvasie@yahoo.com

Abstract. *The software Laça-Palavras (LP) assists in the description of Brazilian Portuguese, by interfacing with files from other programs, especially with the ones of the CLAN system of CHILDES Project, to which this study is aggregate. The interface occurs at two levels: manipulation (data recovery and filtering), and editing (adding and deleting data, and corrections). In interfacing with other applications, the LP creates a data line for automatic phonological transcription, and one for automatic apprehension of the morphology of verbs. This article presents the LP, the syntactic categories that the software is currently working with, and some analysis of data obtained from the file pau003, found in adults' and child's speech registers.*

Resumo. *O aplicativo Laça-palavras (LP) auxilia na descrição do português brasileiro, mediante interface com arquivos de outros programas, sobretudo os do sistema CLAN, da Plataforma CHILDES, ao qual este estudo se agrega. A interface ocorre em dois níveis: manipulação (resgate e filtragem de dados) e edição (acréscimo e exclusão de dados e correções). Ainda em interface com outros aplicativos, o LP cria as linhas de novos dados %pho – para transcrição fonológica automática – e %mor – para depreensão automática da morfologia dos verbos. Apresentam-se o LP, as classes sintáticas com que o programa trabalha atualmente, e uma análise preliminar de dados extraídos do arquivo pau003, relativos aos registros do adulto e da criança.*

1. Introdução

Os computadores e a lingüística computacional tornaram possíveis a catalogação e a análise de quantidade nunca antes conhecida de dados da comunicação verbal, em tempo muito menor. Isso possibilita descrições, comparações e generalizações a partir de uma massa de dados muito mais robusta.

Nessa área, a plataforma CHILDES disponibiliza arquivos e programas para estudiosos do mundo todo. Este artigo apresenta um *software* desenvolvido para pesquisa com o português brasileiro, e mostra alguns resultados relativos à fase atual de um projeto maior – agregado ao CHILDES – do qual este estudo faz parte. Trata-se do projeto Análise Morfológica Automática do PB, cujo objetivo maior é depreender uma gramática automática do português brasileiro. As fases aqui descritas integram-se

Anais da II Jornada de Descrição do Português, páginas 39–46, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

totalmente à lingüística computacional. Nesse sentido, esta pesquisa se aplica a *corpora* de fala, portanto, orais, transcritos segundo o formato da plataforma CHILDES.

2. Plataforma CHILDES

Os avanços tecnológicos ocorridos nos anos 80 possibilitaram a utilização das novas tecnologias disponíveis para automatizar um banco físico de intercâmbio de dados lingüísticos. Assim, em 1984, o Projeto CHILDES (Child Language Data Exchange System) [MacWhinney 2000] foi oficialmente concebido.

O projeto disponibiliza uma gama de dados oriundos de diversas línguas do mundo, com áudio e transcrições completas, feitas por pesquisadores, e automatiza a forma de codificação e análise desses dados. Para tal, conta com duas importantes ferramentas eletrônicas de apoio: o CHAT (Codes for Human Analyis of Transcripts) oferece os padrões e normas para a transcrição e codificação dos dados, e, a partir dessas convenções, a segunda ferramenta do CHILDES atua, o programa CLAN (Computerized Language Analysis), que possibilita análises lingüísticas automatizadas de falas espontâneas, transcritas no CHAT. O CLAN permite, por exemplo, o cálculo automático da Extensão Média do Enunciado (EME, em inglês, MLU – Mean Length Utterance) de um *corpus*, contagens de frequências e análises morfológicas, sintáticas e semânticas [MacWhinney 2011].

Uma vez que as diversas línguas do mundo são morfologicamente distintas, o CLAN conta com uma ferramenta para fazer análises morfológicas: o programa MOR. Cada língua tem seu próprio programa MOR, o qual contém as características morfológicas específicas dessa língua. Um dos objetivos do Grupo de Pesquisa Produtividade Lingüística Emergente (CNPq, certificado pela UFSC), ao almejar depreender uma gramática automática do PB, é criar o programa MOR para o português brasileiro.

A base de dados da plataforma CHILDES contém 44 milhões de palavras faladas em 28 línguas diferentes. É o maior *corpus* de fala atualmente. Já foram construídas gramáticas para 10 línguas, das quais servem de modelo para o português as do italiano e do espanhol.

3. Descrição da Pesquisa

A pesquisa a que este artigo se refere visa à elaboração de regras específicas para a depreensão da referida gramática automática do sistema verbal do português brasileiro, e se insere em um projeto maior, que pretende abranger a morfologia completa do português brasileiro – que é uma língua bastante flexiva – e analisar a fala dirigida à criança.

No nível de desenvolvimento em que se encontra, uma das metas da pesquisa é criar um arquivo com as formas irregulares, particularmente dos verbos, e um arquivo específico com as ultrageneralizações da criança [Marcus *et al.* 1992]. Esses autores, a partir do exame da produção espontânea por 83 crianças do acervo do projeto CHILDES, constataram que a porcentagem de ultrageneralizações é muito baixa e não restrita a uma só fase de desenvolvimento. Por outro lado, os verbos irregulares e de alta frequência são bastante resistentes à ultrageneralização. Finalmente, existe uma fase inicial que os autores denominaram de erros livres, que não resultam da ultrageneralização.

3.1. O *Corpus*

O *corpus* de trabalho foi incorporado ao banco de dados do Projeto CHILDES em 1993, e refere-se à terceira fase do sujeito Pá – criança que está adquirindo o português brasileiro. O arquivo correspondente contém mais de 8.500 enunciados, e registra a época em que a criança estava com 26 meses e 8 dias. O *corpus* é formado pela transcrição de diálogos de três adultos (e outros esporádicos) com a criança, e traz a transcrição fonética somente da fala da criança. Os dados foram transcritos de acordo com o formato CHAT, que é compatível com o programa CLAN. O *corpus* está disponível para os interessados em: http:/childes.psy.cmu.edu/data/Romance/Portuguese/florianopolis.zip, e tem sido usado por vários pesquisadores.

Em outros trabalhos relativos às outras duas fases do sujeito Pá, discutiu-se a emergência das categorias verbais na fase inicial em que ele está adquirindo o português brasileiro [Scliar-Cabral 2007], bem como a codificação da morfologia do PB e a análise da fala dirigida à criança [Scliar-Cabral 2008].

4. O Programa Laça-Palavras

O decorrer da pesquisa exigiu flexibilidade dos dados maior do que a oferecida pelo programa CLAN. Houve a necessidade de se disporem os dados em diferentes formas, bem como de se extraírem deles informações relevantes para a pesquisa, que não eram possibilitadas pelo CLAN. Assim, a fim de acompanhar as etapas do projeto, criou-se um programa para promover interface com os arquivos do CLAN, o Laça-palavras (Vasilévski e Araújo 2011) – para também auxiliar o trabalho dos bolsistas do projeto – o qual trabalha com arquivos em português e está em constante evolução. Desse modo, o LP trabalha em conjunto com o CLAN, mas também disponibiliza recursos próprios, os quais aumentam conforme a programação avança e novas demandas surgem.

4.1. Diretrizes da Interação

Tal interface ocorre em dois níveis: manipulação e interferência. No nível de manipulação de conteúdo, o programa carrega arquivos do CLAN e redistribui seus dados para visualização, permite seleção de conteúdo e gera relatório estatístico, sem alterar, no entanto, o arquivo original. Já no nível de interferência, as ações feitas no Laça-palavras modificam – edição, inserção e exclusão de dados – o arquivo original, mas não se permitem alterações que o CLAN não reconheceria.

Embora tenha sido criado para promover interface com os arquivos do CLAN e interagir, sobretudo, com ele, o LP poderá ser adaptado para ler e manipular arquivos em formato txt, o que ampliará seu alcance e uso. Essa é uma perspectiva para um futuro não tão longínquo. Busca-se, em sua construção, que o Laça-palavras seja interativo, de modo que o usuário consiga usá-lo facilmente.

4.2. Ações Implementadas

No nível de interferência, é possível criar no *corpus* uma linha denominada %pho, para fazer a transcrição fonológica automática, com marcação das sílabas tônicas, de determinado enunciado do arquivo. Essa ação tornou-se possível mediante outra interface criada, dessa vez, entre o Laça-palavras e o tradutor fonológico automático Nhenhém [Vasilévski 2008], o que aumenta o potencial do LP como programa de

interface. O programa permite, inclusive, ajuste da transcrição fonológica para a fonética. Apresenta-se o enunciado da linha 113 do *corpus*, e a respectiva resposta – análise fonológica automática – do Laça-palavras:

*MOT:vamos@va guardar@v esse ?
pho% /'vãmuS gwaR'daR 'esi /

Além disso, é possível criar uma linha para tradução morfológica automática dos verbos marcados, chamada %mor, para a qual foi desenvolvido um algoritmo específico que contém as regras das três conjugações verbais, em seus respectivos modos e tempos [Vasilévski, 2011b] – esse algoritmo ainda está em fase de construção, para que dê conta dos verbos irregulares do PB, com base em Mattoso Câmara Jr. [1997]. Considera-se o tema do infinitivo a forma básica do verbo regular. Apresentam-se os enunciados das linhas 88 e 5651 do *corpus*, e as respostas – análise morfológica automática – do Laça-palavras a elas. Na linha 5651, mostra-se apenas a resposta à forma verbal "fazendo".

*CHI: (a)cabo(u)@v.
V(Inf) RAD+VT SMT SNP PART GER
acabar acaba Ø u
mor%: V|acaba&PPI&3S =finished=

*ISI: que que está@va fazendo@vi lá ?
V(Inf) RAD+VT SMT SNP PART GER
fazer faze -ndo
mor%: V|faze&Ger =doing=

O programa facilita a classificação dos registros, pois disponibiliza, para cada linha pesquisada, a classificação em fala entre adultos [adad], fala de adulto com a criança [adch] e fala da criança [ch]. No que tange a isso, ressalta-se que o processamento automático das unidades morfológicas do PB dos enunciados dos adultos coloca à disposição dos pesquisadores que trabalham com a morfologia do português uma ferramenta poderosa para análises quantitativas e qualitativas. No plano teórico, contribui em nível explicativo para melhor compreensão da construção das gramáticas do PB, particularmente, do sistema verbal, e amplia o entendimento sobre o papel do *input* na construção de tais gramáticas [Scliar-Cabral 2008], além de demonstrar a intuição do adulto, ao utilizar um registro adequado ao nível da criança.

Também, é possível pesquisar mais de uma palavra por vez – mediante o uso da barra reta ou *pipe* (|) entre as palavras –, com a qual se pesquisam e agrupam variações de uma mesma palavra, por exemplo, conforme as notações que ela recebe no *corpus*. Assim, podem-se pesquisar os grupos: vamos|vamo(s) e vai|vamos|vou|fui|foram e se aplicar estatística a eles.

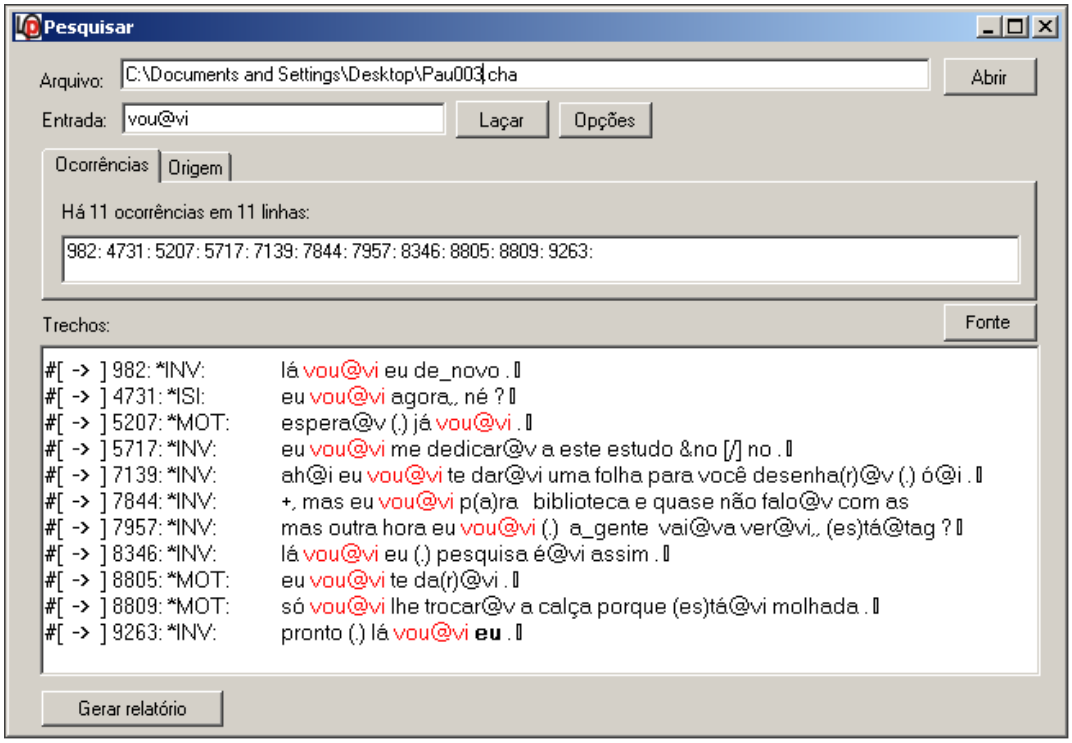

Figura 1. Tela de pesquisa do programa Laça-palavras

O uso e a necessidade têm guiado as ações que devem ser incrementadas no Laça-palavras. Os criadores do programa CLAN estão cientes do Laça-palavras e manifestaram interesse em conhecê-lo e disponibilizá-lo na plataforma CHILDES.

4.3. Classes Sintáticas do Laça-Palavras

Para que o LP atendesse às exigências do projeto na depreensão da gramática automática, definiram-se as classes de trabalho e a nomenclatura a ser utilizada para elas e, após isso, inseriu-se tudo no LP. Vale lembrar que pode haver alguma alteração na nomenclatura, tendo em vista que o programa ainda está em desenvolvimento. Esse incremento amplia sobremaneira as possibilidades de pesquisa no LP.

Para a pesquisa, mostrou-se relevante anotar os verbos diretamente no *corpus*, na linha do enunciado, no sistema CLAN, com @v (verbos regulares – *default*), @vi (verbos irregulares) e @va (verbos auxiliares) – para facilitar, no Laça-palavras, a pesquisa (resgate e filtragem de dados), a análise morfológica automática e, conseqüentemente, a criação da linha %mor. No entanto, para fins de clareza e limpeza do texto, esses símbolos podem ser omitidos na pesquisa a ser feita pelo LP, a critério do usuário. Cabe esclarecer que todos os verbos auxiliares são irregulares, mas a decisão de assinalá-los separadamente se deve ao fato de preparar a computação, posteriormente, das locuções verbais e dos tempos compostos.

A respeito da classificação cabem algumas considerações. Usou-se o termo determinante (det) para os artigos e pronomes adjetivos, critério adotado pela plataforma CHILDES. Do mesmo modo, todos os pronomes substantivos são pro. Sabe-se da dificuldade que a delimitação das locuções adverbiais implica. Uma solução para equacionar esse difícil problema é aplicar o teste da impossibilidade de separá-los pela

interpolação de outra palavra. Convencionou-se unir por _ (*underscore*) os termos da locução no *corpus*.

Até o momento, inseriram-se no Laça-palavras 13 classes sintáticas, e quatro delas têm subclasses – totalizando 21 subcategorias, até o momento. A todas elas associou-se um código, para facilitar a anotação do *corpus*, o resgate e a manipulação de dados, bem como a geração de relatórios estatísticos. As classes são:

Tabela 1. Classes e subclasses sintáticas do Laça-palavras

Classe	Código	Subclasses	Código
Advérbios interrogativos	adv-int	-	-
Artigos	det-art	-	-
Pronomes		adjetivos demonstrativos	det-dem
		adjetivos indefinidos	det-indef
		adjetivos interrogativos	det-int
		adjetivos possessivos	det-poss
		substantivos demonstrativos	pro-dem
		substantivos indefinidos	pro-indef
		substantivos interrogativos	pro-int
		pessoais	pro-pers
		substantivos possessivos	pro-pos
		relativos	pro-rel
Preposições	prep	-	-
Preposições + determinativos	prep-det	-	-
Preposições + pronomes substantivos demonstrativos	prep-pro		-
Conjunções	conj	coordenativas	conj-coor
		subordinativas	conj-sub
Locuções adverbiais	loc-adv	-	
Substantivos	subs	comuns	subs-com
		próprios	subs-prop
Interjeições	int	-	-
Verbos	v	regulares	v
		irregulares	vi
		auxiliares	va
		gerúndio	ger
		particípio	part
		infinitivo impessoal	inf
		infinitivo pessoal	infp
Wordplay	wp	-	-
Singing	s	-	-

Os pronomes pessoais (pro-pers) contêm subclasses que ainda não foram concluídas, como clíticos e subjetivos. *Wordplay* se refere ao uso lúdico dos sons articulados e *Singing* remete à situação em que se canta uma música.

5. Resultados Preliminares

Análises qualitativas, específicas da fala da criança e dos adultos, estão em andamento. Pesquisa quantitativa com o *corpus* anotado mostra que, do total de verbos registrados no *corpus* – 4797, por enquanto, considerando-se a fala dos adultos e da criança – em torno de 50% dos usos são de verbos irregulares, 14% são relativos a verbos auxiliares e 36% são usos de verbos irregulares. Esses dados serão comparados com os achados de Marcus *et al*. (1992). Também, uma primeira comparação demonstrou que a forma "vou" – sem distinguir os casos de @vi e @va – aparece em quantidades muito próximas tanto na fala dos adultos quanto na fala da criança. Essa informação merece análise apurada.

Pesquisas com a classe lexical Pronomes Interrogativos (pro-int) cruzada com o direcionamento do registro dos participantes – [adad], [adch] e [chi] –, mostraram resultados relevantes para os estudos de aquisição da linguagem pela criança, os quais se expõem aqui. O Laça-palavras forneceu 254 itens (100%) para trabalho. Comparou-se o uso dos pronomes interrogativos pelos adultos quando se dirigem à criança e pela criança e quando falam entre si. Exposta a 219 pronomes interrogativos (o_que, do_que, que, quem, qual), pronunciados por adultos, a criança produz 35 (14%). Ela deixa de produzir apenas "do_que", ao qual é exposta três vezes. O pronome mais produzido por ela é "que", 19 vezes, seguido por "quem", nove vezes. Quando os adultos falam entre si, produzem "que" apenas duas vezes, mas, ao falar com a criança, 67 vezes. Quando falam entre si, os adultos produzem 16 (6%) pronomes interrogativos, mas ao falar com a criança, 203 (86%).

A partir disso, verifica-se que a fala dos adultos, quando dirigida à criança, contém bastantes indagações, que revelam estímulos para que a criança fale. Ainda, esse resultado aponta que, nessa faixa etária, a criança está apta lingüística e cognitivamente para processar e produzir a maioria dos pronomes interrogativos, bem como já os compreende perfeitamente [Vasilévski, 2011a].

Vê-se que esses resultados são relevantes para o estudo da fala dirigida à criança e da aquisição da linguagem.

6. Indicativos e Perspectivas

No que tange à aquisição da língua materna, a elaboração do programa Laça-palavras, para auxílio no processamento automático da morfologia, em parceria com o programa CLAN do projeto CHILDES, permitirá não só verificar o quanto a intuição dos adultos permite a adequação de seu registro ao desenvolvimento cognitivo e linguístico da criança e a possível influência da gramática do adulto na construção da gramática da criança, como também colocará à disposição de qualquer pesquisador que deseje investigar o PB um instrumento que lhe facilitará a análise linguística automática de enunciados.

O Laça-palavras e suas interfaces mostram que é possível e necessário ampliar as formas de visualizar e editar os dados, bem como ler outros formatos de arquivo além

do CHAT, a fim de agilizar a descrição do português brasileiro e possibilitar o cotejo dos resultados com outros *corpora*. Isso torna a pesquisa mais consistente e confiável.

O LP facilita o trabalho da equipe do projeto, que se beneficia de seu uso, em paralelo com o programa CLAN. Vislumbra-se um caminho promissor para o programa, pois, apesar do que foi apresentado aqui e de que nem tudo o que o LP faz tenha sido mostrado, o *software* ainda está em fase inicial de desenvolvimento.

Referências

MacWhinney, B. (2011) "Enriching CHILDES for Morphosyntactic Analysis", http://childes.psy.cmu.edu/morgrams/morphosyntax.doc, March.

MacWhinney, B. (2000), The CHILDES Project: Transcription on Format and Programs, New Jersey, Lawrence Erlbaum, 3rd edition, v. I and v.II.

Marcus, G., Ullman, M., Pinker, S., Hollander, M., Rosen, T., and Xu, F. (1992). Overregularization in language acquisition. In *Monographs of the Society for Research in Child Development*, pages 1-182.

Mattoso Câmara Jr., J. (1997), Estrutura da Língua Portuguesa, Vozes, 26. edição.

Scliar-Cabral, L. (2011). Análise Automática da Morfologia do PB (Plataforma CHILDES): aquisição da morfologia verbal. In *VII Congresso Internacional da Abralin*, Curitiba.

Scliar-Cabral, L. (2008). Codificação da Morfologia do PB e Análise da Fala Dirigida à Criança. In *Fórum Lingüístico*, p.69-82. Florianópolis, 5(2).

Scliar-Cabral, L. (2007). Emergência gradual das categorias verbais no Português brasileiro. In *Alfa*, p.223-234. São Paulo, 51(1).

Vasilévski, V. (2011a). Diferenças entre Input e Intake: Evidências na Aquisição de Pronomes Interrogativos. In *Simpósio Internacional Linguagens e Culturas: Homenagem aos 40 anos dos programas de Pós-graduação em Lingüística, Literatura e Inglês da UFSC*, Florianópolis.

Vasilévski, V. (2011b). Programa para processamento automático das unidades verbais do PB. In Análise automática da morfologia do PB (Plataforma CHILDES): aquisição da morfologia verbal, *VII Congresso Internacional da Abralin*, Curitiba.

Vasilévski, V. (2008), *Construção de um programa computacional para suporte à pesquisa em fonologia do português do Brasil*. Tese de doutorado, Florianópolis: UFSC.

Vasilévski, V. e Araújo, M. J. (2011) "Laça-palavras: sistema eletrônico para descrição do português brasileiro" LAPLE-UFSC, v.2010-2011, Florianópolis, https://sites.google.com/site/sisnhenhem/