ANÁLISE AUTOMÁTICA DA MORFOLOGIA VERBAL DO PB: PLATAFORMA CHILDES

Leonor SCLIAR-CABRAL
Universidade Federal de Santa Catarina (UFSC)/CNPq

Vera VASILÉVSKI
(PNPD CAPES)

RESUMO

Os procedimentos de montagem das regras que compõem o aparato para a análise automática da morfologia verbal do PB, dentro da plataforma CLAN, serão apresentados e debatidos. Comparando-se a formalização das classes sintáticas e respectivas regras do espanhol e do italiano com as do PB, chegou-se à conclusão de que elas deveriam ser reformuladas, particularmente no que diz respeito à análise automática dos morfemas verbais. Tendo em vista o nível de previsibilidade dos morfemas do sistema de verbos do PB, apresenta-se a formalização das regras morfológicas que compõem esse sistema para os verbos regulares das três conjugações e a formalização em algoritmo, bem como o trabalho que a antecedeu na programação informatizada que identifica automaticamente as formas verbais do português, classificando-as segundo modo/tempo, pessoa/número, em compatibilidade com o sistema CLAN, da plataforma CHILDES (MacWHINNEY, 2000, 2008). Discutem-se as dificuldades encontradas na conversão e as decisões que foram tomadas para superá-las e mostra-se a criação automática de uma linha fonológica. Este trabalho é realizado com apoio do CNPq e da CAPES, entidade do governo brasileiro voltada para a formação de recursos humanos.

1. Histórico

O Grupo Integrado Produtividade Linguística Emergente do CNPq há anos vem alimentando o maior banco mundial de dados de linguagem verbal, a plataforma CHILDES, conforme pode ser visualizado e ouvido no site: http:/childes.psy.cmu.edu/data/Romance/Portuguese/florianopolis.zip, uma vez que todos os enunciados, tanto dos adultos quanto os da criança são seguidos de *bullets* que, quando clicados permitem sua audição. Há três corpora, correspondentes à fase 1 (20m e 21d), à fase 2 (22m e 20d) e à fase 3 (26m e 08d) do sujeito Pá, cujos enunciados também foram transcritos foneticamente (*broad transcription*). O principal achado da pesquisa foi considerar o acento de intensidade como morfema verbal (suprafixo), com a função de assinalar na 1ª fase diferenças aspectuais (posteriormente, redundante e cumulativamente também assinalará tempo/modo). Propusemos, então a implementação[1] da fórmula de Mattoso Camara Jr. (2004:134) que passa a: **T(R+VT) + SF (SMTA+SNP +SPF)**.

O mundo contemporâneo dos computadores e da linguística computacional tornou possível a catalogação e análise de uma quantidade antes nunca conhecida de dados da comunicação verbal, em tempo muito menor. Isto possibilita comparações e generalizações a partir de uma massa de dados muitíssimo mais robusta. A base de dados da plataforma CHILDES, com a qual o presente projeto opera, contém 44 milhões de palavras faladas em 28 línguas diferentes. Trata-se do maior *corpus* de fala atualmente existente. Em segundo lugar, vem o *British National Corpus*, com 5 milhões de palavras.

Todos os dados do sistema CHILDES estão codificados de forma consistente num formato de transcrição denominado CHAT, inclusive os dados da 3ª fase do sujeito Pá, dos quais será depreendida a gramática automática. Atualmente já foram construídas gramáticas MOR de 10 línguas: cantonês, holandês, inglês, francês, alemão, hebraico, japonês, italiano e espanhol, das quais servirão de modelo para a depreensão da gramática do PB as gramáticas do italiano e do espanhol e as de

linguistas brasileiros (Bechara (1999); Borba *et al.* (2002); Castilho (1989, 2002a e b); Cunha & Lindley-Cintra (1987); Ilari (2002); Ilari & Basso ___(2006); Kato (2002); Koch (2002); Mattos e Silva (2001); Moura Neves (2000, 1999); Naro; Scherre (1993); Preti (1993); Roncarati & Abraçado (2003)).[1]

O projeto ora apresentado tem como principal meta colocar à disposição dos pesquisadores uma ferramenta que lhes possibilite a análise morfológica automática dos enunciados que constituem os *corpora* coletados do PB.

O Grupo Integrado do CNPq, Produtividade Linguística Emergente, já realizou a análise morfológica manual dos enunciados da criança, nas fases 1 e 2, exemplificados a seguir:

Fase 1 (20m e 21d):

47 *CHI: <não é> [>]!
48 %pho: ˊnə̃w ˊɛ
49 %mor: neg|não=not v:cop1|s-TV2&IPFVM1=is!

A linha 47 é a linha principal (*main line*), contendo um enunciado com dois itens da criança (CHI); a linha 48 %pho é a transcrição fonética e a linha 49 %mor é a análise morfológica manual, em que cop1 é a cópula 1, cujo radical do verbo ser é s-, com a vogal temática da 2ª conjugação, no imperfectivo (as pessoas do discurso ainda não estão gramaticalmente assinaladas).

Fase 2 (22m e 20d):

52 *CHI: cadê ota [= outra] cadeira?
53 %pho: kaˊde ˊota kaˊdeɾɐ
54 %mor: wh:proloc|cadê=where det|ota=another n|cadeira=chair?

[1] Para um detalhamento, consulte-se Scliar-Cabral (2007).

Uma formalização semelhante, porém, expressando a gramática do PB, será o *output* na linha %mor, quando for disponibilizada a gramática automática para análise de *corpora* formatados de acordo com o formato CHAT.

2. Codificação dos paradigmas das classes sintáticas

Para a preparação da gramática automática, cujas regras e respectivos algoritmos dos tempos simples dos verbos regulares serão explicados nesse artigo, já foram codificados os paradigmas das classes sintáticas, a seguir exemplificados:

Advérbios interrogativos
onde {[scat adv:int]} =where=

Artigos
a {[scat art]} "o&FEM&SG" =the=

Pronomes adjetivos demonstrativos
aquela {[scat det:dem] "aquele&FEM&SG" =that=

Pronomes adjetivos indefinidos
algum {[scat det:indef]} "algum&MASC&SG"=some=

Pronomes adjetivos interrogativos
que {[scat det:int]} =what=

Pronomes adjetivos possessivos
meu {[scat det:poss]} "meu-1S&MASC&SG" =my=

Pronomes substantivos demonstrativos
a {[scat pro:dem] "o&FEM&SG" =in English it is included in wh form=

Pronomes substantivos indefinidos
algo {[scat pro:indef]} =something=

Pronomes substantivos interrogativos
o_que {[scat pro:int]} =what=

Pronomes pessoais
% subject case
eu {[scat pro:pers]} "eu&1S&SUBJ" =I=
% forms that are the same as subject and object (in the last case, always preceded by preposition)
você {[scat pro:pers]}"você&2S&SG&OBJ" =you=
% clitics
me {[scat pro:pers]} "eu&1S&OBJ" =me=

Pronomes substantivos possessivos
minha {[scat pro:pos]} "meu-1S&FEM&SG" =mine=

Pronome relativo
quem {[scat pro:rel]} =who=

Preposições
a {[scat prep]} =to=

Preposições + determinativos
à {[scat prep]} «a~det:art|o&FEM&SG» =to the=

Preposições + pronomes substantivos demonstrativos
à {[scat prep]} «a~pro:dem|o&FEM&SG» =to the one=

Conjunções coordenativas
contudo {[scat conj:coor]} =nevertheless=

Conjunções subordinativas
antes_que {[scat conj:sub]} =before=

Substantivos comuns
adultos {[scat n][gen masc]} "adulto-PL" =adults=

Substantivos próprios
Ana {[scat n][gen fem]}

Uma solução para equacionar o difícil problema da delimitação das locuções, cujos termos vêm ligados por _, é aplicar o teste da impossibilidade de separá-los pela interpolação de outra palavra. O critério não foi aplicado para os tempos compostos e locuções verbais, uma vez que seus respectivos auxiliares são arrolados em paradigma específico. Veja-se, a seguir, um exemplo de codificação de locução adverbial listada no *corpus* PAU003:

Locuções adverbiais
ao_mesmo_tempo {[scatadv:loc]} =at_the_same_time=

A seguir serão apresentadas as ferramentas produzidas por Vera Vasilévski, utilizadas pelos pesquisadores do Grupo Integrado Produtividade Linguística Emergente, como auxiliares na depreensão da gramática automática do PB.

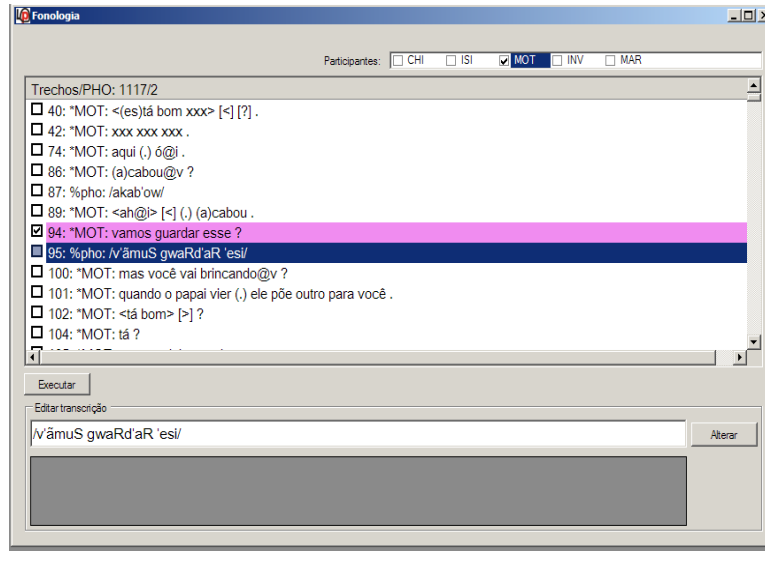
3. Trajetória

Em 2008, como resultado da tese de doutorado *Criação do Sistema de Conversão Grafema-fonema Nhenhém*, foi criado o programa de conversão grafêmico-fonológica automática ***Nhenhém*** e, em 2009, a primeira atualização. Em 2010 ocorreu a conversão do Nhenhém para outra linguagem de programação, a criação do Nhenhém silabador, a versão atualizada do Nhenhém silabador (entrada da morfologia, na separação silábica de palavras compostas por justaposição) e a Interface entre o Nhenhém e os arquivos Clan.

A interface entre o programa Clan foi feita com um programa específico criado para auxiliar o trabalho dos bolsistas do projeto. Chama-se LAÇA-PALAVRAS e abriga os demais programas e funções, além de ler os arquivos do Clan.

A interface ocorre em dois níveis: manipulação de conteúdo (lê os dados e os dispõe em estatística, sem alterar o arquivo original) ou interferência nos arquivos (modifica/edita-os) e, cria, simultaneamente a linha %pho, conforme o quadro abaixo:

QUADRO 1: Conversão grafo-fonológica simultânea da linha 94.



4. Regras de alomorfia das VTs, dos SMT e SNP dos verbos regulares

Como primeiro passo, foram formalizadas as regras de alomorfia das vogais temáticas (VT) das três conjugações dos verbos regulares, cujas ocorrências são marcadas na linha principal com @v, como, por exemplo, na linha 48 *CHI: acende@v a luz. Tais regras servem de base para o algoritmo para sua inserção no programa.

QUADRO 2: Exemplo de formalização das regras para a vogal temática |-a-| da 1ª conjugação.

VT	se reescreve	como	em contexto	Exemplos
|-a-|	→	Ø	__õ#	cantØo
			__e	cantØe, cantØes
		á	´__ CV.CVC	cantávamos
			´__ CCV.CVC	cantássemos
			´__ CVCC	cantáreis
		ã	(VC)C´__o#	dão, estão
		e	__i#	cantei
		o	__u#	cantou
		a	...	cantamos

A seguir, passou-se à formalização das regras de alomorfia do sufixo modo-temporal (SMT), conforme exemplo no quadro 3.

QUADRO 3: Formalização da regra para o sufixo modo (-temporal) (SM(T)) do gerúndio

SMT
|-ndo|

O presente do indicativo, como tempo primitivo, apresenta Æ para SMT.

QUADRO 4: Formalização da regra para o sufixo modo-temporal (SMT) do Pretérito imperfeito do Indicativo.

SMT	se reescreve	como	em contexto	Exemplos
|-va-|	→	-ve-	a __ i	cantáveis
		-va-	...	cantava, cantávamos

O pretérito perfeito do indicativo, como tempo primitivo, apresenta Æ para SMT, com exceção da 3ª pessoa do plural, -ra-, que é uma forma marcada.

QUADRO 5: Formalização da regra para o sufixo modo-temporal (SMT) do Futuro do_presente do Indicativo.

Futuro do presente

SMT	se reescreve	como	em contexto	Exemplos
|-rá-|	→	-re- -rã- -rá	/ { _i, _m } _o# ...	cantareis cantaremos cantarão cantarás

Finalmente, foram formalizadas as regras de alomorfia para o sufixo número-pessoal (SNP), das quais damos dois exemplos:

QUADRO 6: Formalização da regra para o sufixo número-pessoal (SNP) de 1ª pessoa do plural (só apresenta alomorfe em juntura fechada com os pronomes pessoais clíticos).

1ª. pessoa do plural

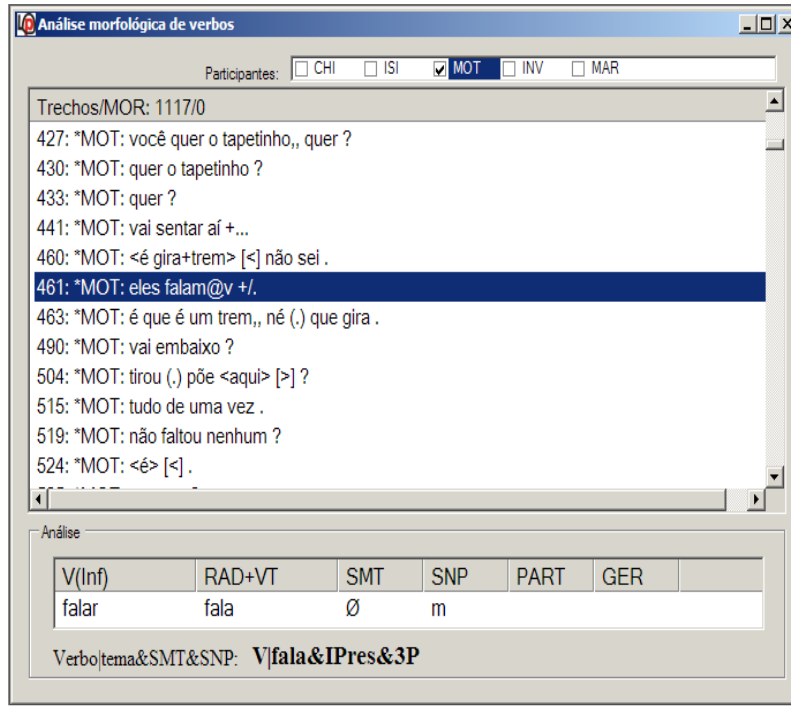
SNP
|-mos|

QUADRO 7: Formalização da regra para o sufixo número-pessoal (SNP) de 3ª pessoa do plural.

3ª. pessoa do plural (r)

SNP	se reescreve	como	em contexto	Exemplos
|-m#|	→	-o# / -m#	/ ‘ã __ # / ...‘#	estão, cantarão cantam, falavam

QUADRO 8: Exemplo de análise morfológica automática da ocorrência “falam” para inserção na linha %MOR)



Resultados

Conforme se pode depreender, encontra-se bastante adiantado o projeto de criação do programa que analisará automaticamente a morfologia do PB. Nesse ínterim, codificaram-se os paradigmas de quase todas as classes sintáticas, elaboraram-se as regras alomórficas das vogais temáticas e dos sufixos modo-temporais e número-pessoais do sistema escrito dos verbos regulares do PB, bem como foram construídas poderosas ferramentas de investigação como o programa Nhenhém (em várias versões), o silabador e o Laça-palavras, além dos algoritmos de conversão para linguagem de máquina, mas ainda teremos de resolver conflitos decorrentes das ambiguidades por meio de regras específicas.

Referências

BECHARA, E. **Moderna gramática portuguesa**. 37ª ed. ver. e amp. Rio de Janeiro: Lucerna, 1999.

BORBA, F. S. *et al.* **Dicionário de usos do português do Brasil**. São Paulo: Ática, 2002.

CASTILHO, A. T. (Org.). **Português culto falado no Brasil**. Campinas: UNICAMP, 1989.

_____. (Org.). **Gramática do português falado**, 4ª ed. revista. A ordem. Campinas: UNICAMP/FAPESP. v. I, 2002a.

_____. (org.). **Gramática do português falado**, 3ª ed. revista. As abordagens. Campinas: UNICAMP/FAPESP. v. III, 2002a.

CUNHA, C.; LINDLEY-CINTRA, L.F. **Nova gramática do português contemporâneo**. Rio de Janeiro: Nova Fronteira, 1987.

ILARI, R. (Org.). **Gramática do português falado**. 4ª ed. revista. Níveis de análise linguística. Campinas: UNICAMP, v. II, 2002.

_____; BASSO, R. **O português da gente** – a língua que estudamos, a língua que falamos. São Paulo: Contexto, 2006.

KATO, M. (Org**.). Gramática do português falado**. 2ª ed. revista. Convergências. Campinas: UNICAMP, v. V, 2002.

KOCH, I. G. (Org.). **Gramática do português falado**. 2ª ed. revista. Campinas: UNICAMP, v. VI, 2002.

MacWHINNEY, B. **The CHILDES Project:** Transcription on Format and Programs. 3ª ed. New Jersey: Lawrence Erlbaum, 3ª ed, v. I e II, 2000.

_____. **Enriching CHILDES for Morphosyntactic Analysis**. Plataforma CHILDES, 2008. Disponível em: <http://childes.psy.cmu.edu/morgrams/morphosyntax.doc> Acesso em: 22 set. 2008.

MATTOS E SILVA, R. V. (Org.) **Para a história do português brasileiro**. Primeiros estudos. São Paulo: Humanitas/FAPESP, v. II, 2001.

MATTOSO CAMARA JR., J. **Para o estudo descritivo dos verbos irregulares**. *In*: FALCÃO UCHOA, C. E. (org.). Dispersos de J. Mattoso Câmara Jr. Rio de Janeiro: Lucerna, 2004, p.131-146.

MOURA NEVES, M. H. de. (Org.). **Gramática do português falado**. 2ª ed. revista. Novos estudos. Campinas: UNICAMP/Humanitas, v. VII, 1999.

_____. **Gramática de usos do português**. São Paulo: UNESP, 2000.

NARO, A. SCHERRE, M. M. P. **Sobre as origens do português popular do Brasil**. DELTA, v. 9, p. 437-54, 1993.

PRETI, D. (Org.). **Análise de textos orais. São Paulo:** FFLCH/USP, 1993.

RONCARATI, C.; ABRAÇADO, J. (Orgs.). **Português brasileiro:** contato linguístico, heterogeneidade e história. Rio de Janeiro: Letras/ FAPERJ, 2003.

SCLIAR-CABRAL, L. **Emergência gradual das categorias verbais no português brasileiro**. ALFA – Revista de Linguística, 51(1), os. p. 222-234, 2007. Disponível em: <http://www.alfa.ibilce.unesp.br/index.php>

VASILÉVSKI, V. **Construção de um sistema computacional para suporte à pesquisa em fonologia do português do Brasil**, Tese (Doutorado) Pós-Graduação em Linguística, Universidade Federal de Santa Catarina, 2008.

Créditos:

Tradução das glosas para o inglês: Cloves Cardozo (bolsista IC/PIBIC)

Programação: Márcio Araújo (Engenharia Eletrônica) e Vera Vasilévski