AQUISIÇÃO DA LINGUAGEM

VOLUME ESPECIAL - 2012

Preparando a gramática automática do português brasileiro

Leonor Scliar-Cabral (UFSC/CNPq)
Vera Vasilévski (UFSC – PNPD/CAPES)

RESUMO: Define-se o que é a gramática automática, descrevem-se os procedimentos para a preparação da linha principal para a análise morfológica por regras e expõem-se os critérios adotados para alimentar o léxico ao qual são pareados os itens reconhecidos na linha principal. Os referenciais teóricos são o modelo competitivo (MacWHINNEY, 2000a, 2008) e os dos autores brasileiros que se debruçaram sobre o PB atual. O *corpus* originário provém da tese de doutorado de Scliar-Cabral (1976) e os resultados já obtidos são a formatação de todas as linhas do *corpus*, a organização dos léxicos e a elaboração dos algoritmos (VASILÉVSKI; ARAÚJO, 2011, VASILÉVSKI *et al.* 2012).

Palavras-chave: gramática automática; português brasileiro; léxicos; ambiguidade; CHILDES.

Introdução

O objetivo deste artigo é descrever os procedimentos para a preparação da linha principal (*input*), de acordo com a plataforma, Chat com vistas à a análise morfológica por regras gramaticais, para aplicação da gramática automática do PB que está sendo elaborada, da qual resultará a linha %mor (*output)*. Expomos, em adendo, os critérios adotados para alimentar os arquivos correspondentes aos léxicos construídos para as classes gramaticais, aos quais são pareados os itens reconhecidos na linha principal.

A construção da gramática automática do PB teve início com o projeto *Codificação da morfologia do PB e análise da fala dirigida à criança*, que se propôs a adaptar à morfologia do português brasileiro os procedimentos de montagem das regras que compõem o aparato para a análise automática da morfologia, dentro da plataforma Childes. Assim, usa-se o programa Clan (CHILDES, 2012) para rodar o *corpus*, de forma que se respeitem seus

---

princípios de funcionamento e a nomenclatura por ele utilizada, na construção dos algoritmos, na classificação dos dados e na organização dos léxicos, ao mesmo tempo em que se criam formalizações específicas que o português brasileiro exige.

1. Gramáticas automáticas

As gramáticas automáticas reconhecem os dados de fala e de escrita codificados na linha principal, de forma consistente, num formato de transcrição denominado *chat* (inclusive os dados da 3.ª fase do sujeito Pá – pau003.chat), sob padrões determinados, emparelham-nos com os itens classificados nos léxicos e os rotulam em categorias gramaticais, na linha morfológica (%mor). Em adendo, além da linha principal, poderá haver a transcrição fonética (%pho), particularmente da fala da criança e/ou da fala atípica, bem como linhas com explicações (%com, %err), para suprir o contexto situacional e erros.

2. Histórico da gramática automática do PB

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 sítio: <http:/childes.psy.cmu.edu/data/Romance/Portuguese/florianopolis.zip>. 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 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.

Atualmente, já foram construídas gramáticas MOR de dez 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. Comparando-se, porém, a formalização das classes sintáticas e respectivas regras do espanhol e do italiano, que servem de modelo, com as já efetuadas para o PB, chegou-se à conclusão de que elas necessitavam ser refinadas e expandidas, com o concurso de especialista em linguística computacional, com sólida formação em linguística.

Os referenciais teóricos que servem de apoio para a categorização das classes sintáticas do português brasileiro são: Basílio (2004, 2000, 1999, 1987, 1980), Borba *et al*. (2002), Castilho (1989, 2002a, 2002b), Ilari (2002); Ilari e Basso (2006), Moura Neves (2000, 1999). Obviamente, continuamos consultando os clássicos de Mattoso Câmara Jr. (1971; 1976; 1997 [1970]; 2004). Propusemo-nos, pois, colocar à disposição da comunidade científica um programa que possibilite o processamento automático das unidades morfológicas do português brasileiro (PB) pelo programa CLAN do projeto CHILDES (MacWHINNEY, 2000b), porém, com as adaptações e expansões acima mencionadas. O *corpus* é constituído dos enunciados da 3.ª fase de PAU (26 meses e 08 dias), já transcritos canonicamente, com os respectivos *bullets* e transcrição fonética, na linha %pho, e dos enunciados dos adultos, produzidos pelos pais da criança, ambos falantes da variedade paulista (cidades de Campinas e São Paulo), o pai, linguista, com doutorado, e a mãe, psicóloga; pela investigadora, falante da variedade gaúcha (cidade de Porto Alegre), na época, em fase de doutorado em Linguística e, esporadicamente, sobre alguns enunciados das empregadas.

Dispomos, hoje, de todos os arquivos com as classes sintáticas (léxicos), que permitem o pareamento com a palavra reconhecida na linha principal. Por exemplo, reconhecidas as palavras "seu" e "quarto" na linha principal (a linha onde estão as três letras em maiúsculas que designam os participantes, no exemplo abaixo), o programa as emparelha com os respectivos itens no léxico, já assinalados com as categorias gramaticais, de que resulta o *output* na linha %mor:

@Situation: na sala, brincando
*MOT: vem@vi (.) (es)pera@v lá no seu quarto pra@p [: para] Leonor ver@vi seus brinquedos ?
*INV: vamo(s)@vi [: x2] lá [: x2] .
*MOT: vamos@vi lá no seu quarto (.) vamos@tag ?
*CHI: lã [: lá] dento@p [: dentro] .

det-pos.cut
seu {[scat det:pos]} "seu-3S&MASC" =your=
em que det-pos.cut é o nome do arquivo no léxico, ou seja, determinante possessivo; seu é a ocorrência na linha principal; scat det:pos é a categoria sintática, ou seja, determinante possessivo; "seu-3S&MASC&SG" indica os morfemas presos, ou seja, terceira pessoa do singular e gênero masculino e =your= é tradução para o inglês. O *output* da análise morfológica, na linha %mor é:

%mor: det:pos|seu-3S&MASC =your

n-comum.cut
quarto {[scat n][gen masc]} =bedroom=
em que n-comum.cut é o nome do arquivo no léxico, ou seja, nome comum; quarto é a ocorrência na linha principal; [scat n] é a categoria sintática, ou seja, nome comum; [gen masc] é o gênero intrínseco, ou seja, masculino e =bedroom= é tradução para o inglês. O *output* da análise morfológica, na linha %mor, é:

%mor: n|quarto-MASC =bedroom

3. Procedimentos

Conforme mencionado na definição, três tarefas correm paralelas: preparar a linha principal, alimentar os léxicos e construir os algoritmos que gerarão a linha morfológica (%mor). Neste artigo, discutiremos os procedimentos das duas primeiras tarefas.

3.1 Preparando a linha principal: Desafios

A fala não é constituída de estruturas perfeitas, pois faltam, às vezes, constituintes essenciais à frase, como nos exemplos abaixo, em que não consta o núcleo da frase verbal:

*INV: também (.) primavera,, né@tag ?
*MOT: só mais um_pouco (.) Paulo .

*INV: para quê (.) meu filho ?

Também ocorrem deslocamentos agramaticais, como no exemplo a seguir: mais que tudo, pausas plenas ou vazias de processamento interrompem o elo que deveria unir as unidades mais básicas dos constituintes, provocando descontinuidades:

*MOT: do Mickey você não mostrou .
*INV: embora (.) **&é** a_gente tem@va que assinalar@v quantas vezes a criança imita@v,, sabe@tag ?
*INV: como será_que **&é** estragou@v;; hein@tag ?

Frequentemente, o pronome interrogativo átono "que" é reduplicado com a função de redundância, dada sua fraca saliência perceptual, como em:

*INV: que **&que** você faz@vi com +...

Observe que tais elementos (em negrito) são precedidos de &. Isso significa que serão descartados do léxico, e o programa não os computará como classes sintáticas; o terminador +... indica que a frase foi interrompida.

Uma grande dificuldade para perceber os itens produzidos pelos falantes decorre da superposição de vozes (*overlapping*). Isto nos tem suscitado uma indagação teórica, pois refuta o postulado das condições felizes da pragmática, pelo qual cada interlocutor espera sua vez de falar, a qual é sinalizada:

*INV: faz@vi a barba eu quero@va ver@vi aqui como é_que você faz@vi .
*ISI: xxx dinheiro xxx .
*MOT: é@co .
%com: overlapping nos três últimos enunciados=the three last utterances overlapped (Note a forma como é codificada a clivagem: é_que)

O português brasileiro falado está em constante evolução, e tais mudanças se observam, em particular, nos vocábulos átonos, como, por exemplo, na preposição "para", em que ocorre o processo de apócope da vogal /a/: /para/ → /pra/, a qual, no processo de sândi externo, se for seguida do artigo /o/, resultará em /pro/. Como a gramática automática é uma gramática do sistema escrito, que se atrasa em relação às mudanças do sistema oral, para que não se perca a informação de como o enunciado foi produzido, adota-se a seguinte notação: pro@p [: para + o], em que a extensão @p significa mudança fonética estável. O programa emparelha com o que está entre os colchetes. A informação à esquerda é preservada para estudos sobre variação sociolinguística. O mesmo ocorre na transcrição do verbo tá@p [: está@vi]. Na transcrição dos verbos, observe as informações @v, @vi e @va, respectivamente para verbos regulares, verbos irregulares (na verdade, os consideramos verbos com formas irregulares em seu paradigma de conjugação) e verbos auxiliares (todos também possuem formas irregulares em seu paradigma), uma vez que, ao contrário das gramáticas do espanhol e do italiano, o emparelhamento da forma reconhecida na linha principal não se dá com a que consta no paradigma completo do verbo, e sim com o tema e sufixos modo-temporais e número-pessoais, conforme os algoritmos construídos por Vasilévski e Araújo (2011).

Trata-se de uma divergência teórica e metodológica, pois, partindo-se do pressuposto de que os alomorfes verbais escritos nos verbos regulares (e em muitas derivações dos verbos irregulares) são grafológica e gramaticalmente condicionados, portanto, previsíveis por regras; partindo-se igualmente do pressuposto de que no programa MOR há três arquivos: ar.cut (o arquivo das regras alomórficas); o cr.cut (o arquivo das regras de concatenação) e o sf.cut (o arquivo dos marcadores das formas especiais, isto é, formas marcadas), em que ar.cut são regras que geram variantes alomórficas a partir dos radicais ou temas e cr.cut são regras que especificam as combinações possíveis entre os morfemas da esquerda para a direita numa palavra e ainda de que os afixos são listados no arquivo 0affix.cut., pode-se concluir que uma formatação por regras atende aos princípios de simplicidade (*Occam's razor*), generalidade e previsibilidade da filosofia da ciência.

Com efeito, qualquer ocorrência verbal, mesmo que o verbo ainda não esteja consignado no léxico, poderá ser automaticamente analisada. Abaixo, na Figura 1, está a análise morfológica automática da ocorrência entoou@v (após o programa ter tentado emparelhar os infinitivos possíveis (entoar, *entoer e *entoir), somente a forma "entoar" foi encontrada no léxico dos verbos; existe uma regra alomórfica que transforma a vogal temática "a" em "o", antes do contexto final "u"; PPI é o acrônimo de Pretérito Perfeito do Indicativo e 3S é a sigla para terceira pessoa singular):

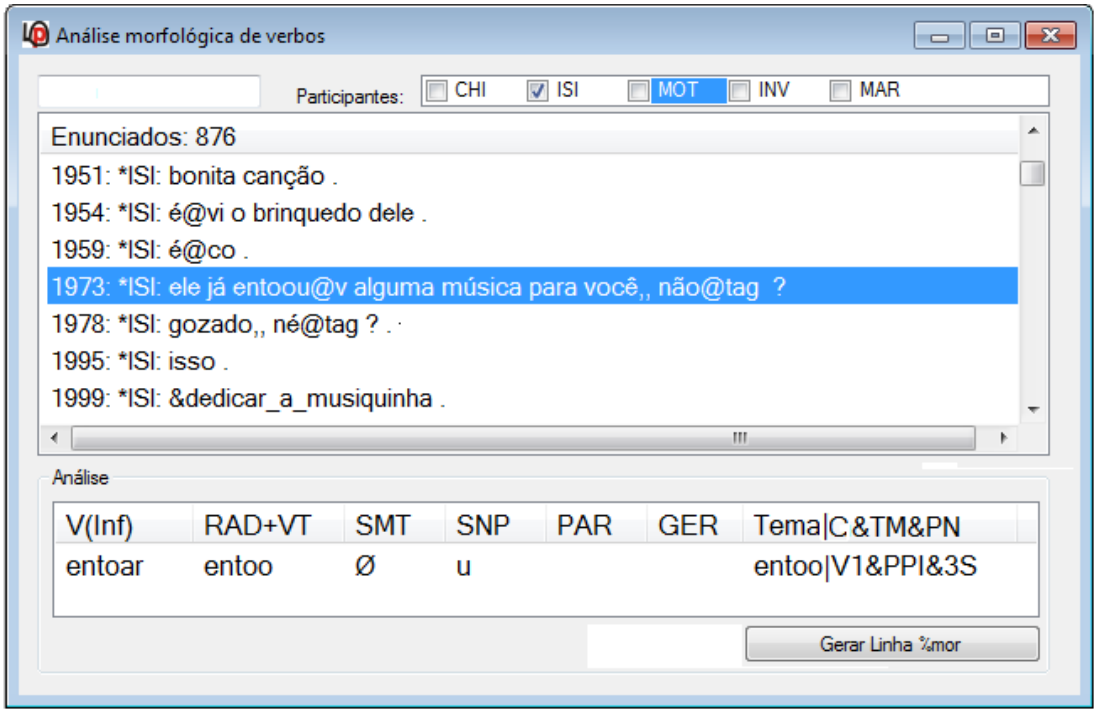

Figura 1: Amostragem do analisador morfológico automático para a ocorrência entoou@v.

Outro desafio à preparação da linha principal consiste no registro dos enunciados da criança. Conforme já explicado, o programa só reconhece as palavras que estejam codificadas no formato de transcrição denominado chat. Como o sistema fonológico da criança ainda dista daquele do adulto e um mesmo item apresenta muitas flutuações, é necessário traduzi-lo para a forma canônica, a fim de ele ser reconhecido, preservando-se, porém, a emissão original, a

fim de que não seja atribuída à criança uma competência linguística que ela ainda não possui. Por outro lado, a preservação do enunciado da criança se presta a análises sobre a aquisição do componente fonológico e respectivas teorias. Sendo assim, na linha principal, à esquerda dos colchetes se escreve a forma produzida ou simplesmente se colocam entre parênteses as unidades faltantes, além de, na linha fonológica (%pho), registrá-la em transcrição fonética, conforme abaixo:

*CHI: vamos@vi depoi(s) .
%pho: ˊvə̃muz deˊpojs
*CHI: acendi@p [: acende(r)@v] .
%pho: ˊvãmuz asẽˊdʒi

3.2. Preparando o léxico: desafios

A preparação do léxico exige que todos os itens lexicais que aparecem na linha principal sejam categorizados em classes sintáticas. Defrontamo-nos com os seguintes desafios:

3.2.1. Ambiguidades

É sabido que um mesmo significante pode servir a várias classes sintáticas, identificadas (não conscientemente) pelo usuário conforme seus privilégios de ocorrência, ou contextos de distribuição. Essa identificação, que ocorre em milissegundos, os programas de computador não conseguem fazer. O exemplo mais crucial, no português brasileiro, é o do significante "que", o qual pode ser pronome interrogativo (substantivo ou adjetivo); pronome relativo; conjunção subordinativa integrante, comparativa, consecutiva, além de integrar várias locuções conjuncionais; conjunção coordenativa explicativa; advérbio de intensidade; preposição; além de entrar nas clivagens.

Utilizamos três critérios para resolver tal impasse: o primeiro consiste em considerar como *default* a classe que apresenta maior número de ocorrências; o segundo consiste em colocar o travessão que o une numa locução, como em o_que, quando for pronome interrogativo, ou em antes_que (locução conjuncional subordinativa final); o terceiro consiste em colocar uma extensão (última alternativa), como @ci, para a conjunção integrante:

*MOT: é_que é@vi um trem,, né@tag (.) que gira@v .

Observe, no exemplo acima, como "o que" é reconhecido na ocorrência é_que, como integrando uma clivagem. Já em:

*INV: tá@p [: está@vi] o_ quê ?

trata-se de um pronome interrogativo. Em:

*MOT: pegou@v (.) &que assim_que chegou@v diz@vi que@ci no_ar &numa uma bactéria uma coisa que .

&que é uma hesitação, sendo descartado; assim_que é uma locução subordinativa conjuncional temporal; que@ci é uma conjunção subordinativa integrante; coisa que é *default* (pronome relativo).

3.2.2. Limites de palavra

Uma das questões mais complexas em linguística foi a delimitação de palavra, a tal ponto que Vendryés, ao encerrar os trabalhos do VI Congresso Internacional de Linguistas, em Paris, assinala o desacerto, em virtude dos critérios distintos, ora focalizando o significante, ora o significado, ora a unidade do signo, ora do ponto de vista paradigmático, ora sintagmático (SCLIAR-CABRAL, 1971, p.149). Assumiu-se como limite de palavra o critério chomskyano de classe sintática, que preenche o constituinte mais baixo na cadeia de derivação. Se houver espaços em branco separando os itens que constituem uma classe sintática, aqueles são substituídos por travessões (exceto nas locuções verbais e tempos compostos, conforme já examinado).

Detalharemos, a seguir, as classes sintáticas contempladas no léxico, inclusive as classes específicas em aquisição da linguagem e aquelas não articuladas sintaticamente, como, por exemplo, as interjeições.

4. O léxico

Para cada classe, separamos alguns exemplos, e daremos algumas explicações de classes não constantes na NGB (1959).

4.1 Classes sintáticas

adj.cut
aberto {[scat adj]} =open=
acostumado {[scat adj]} =accustomed=
ágrio {[scat adj]} =sour=
alto {[scat adj]} =tall=
amadurecida {[scat adj]} "amadurecido&FEM" =matured=

adv-int.cut
aonde {[scat adv:int]} =where_to=
cadê {[scat adv:int]} =where_is=
como {[scat adv:int]} =how=
da_onde {[scat adv:int]} =where_from=
de_onde {[scat adv:int]} =where_from=

adv.cut
agora {[scat adv]} =now=
aí {[scat adv]} =there=
ainda {[scat adv]} =still=

além {[scat adv]} =besides=
ali {[scat adv]} =there=

adv.loc.cut
à_perfeição {[scat adv:loc]} =perfectly=
a_primeira_vez {[scat adv:loc]} =the_first_time=
a_última_vez {[scat adv:loc]} =the_last_time=
ao_fundo {[scat adv:loc]} =background=
ao_mesmo_tempo {[scat adv:loc]} =at_the_same_time=

aum.cut
grandão [{scat adj&aum}] =very_big=

ch-inv.cut (palavras inventadas pela criança)
piti@c [: disco] {[scat n][gen masc]} =record=
Didiz@c [: Luiz]+Gonzaga {[scat n:prop][gen masc]
Mabeti@c [: Margarete] {[scat n:prop][gen fem]}
Mimi@c (obs. reporting a soccer game)
Mindo@c (obs. reporting a soccer game)

cl.cut (clivagens)
é_que {[scat cl]}
foi_que {[scat cl]}
que_é {[scat cl]}
será_que {[scat cl]}
sabe_que {[scat cl]}

co.cut (confirmativos)
ahn {[scat co]} =ahn=
deixa {[scat co]} =let=
é_verdade {[scat co]} =it's_true=
ficou {[scat co]} =it stayed=
foi {[scat co]} =it was=

conj-coor.cut
contudo {[scat conj:coor]} =nevertheless=
e {[scat conj:coor]} =and=
entretanto {[scat conj:coor]} =nevertheless=
logo {[scat conj:coor]} =so=
mas {[scat conj:coor]} =but=

conj-sub.cut
a_não_ser_que {[scat conj:sub]} =only_if=
ainda_que {[scat conj:sub]} =although=
antes_que {[scat conj:sub]} =before=
apesar_de_que {[scat conj:sub]} =although=
assim_que {[scat conj:sub]} =as_soon_as=

cont.cut (continuativos da narração)
então {[scat cont]} =then=
daí {[scat cont]} =then=
mas {[scat cont]} =but=

den.cut (denominadores, vestígios de fase anterior)
esse {[scat.den]}

det-art.cut
a {[scat art]} "o&FEM&SG" =the=
as {[scat art]} "o&FEM&PL" =the=
o {[scat art]} "o&MASC&SG" =the=
os {[scat art]} "o&MASC&PL" =the=
um {[scat art]} "um&MASC&SG" =a=

det-dem.cut
aquela {[scat det:dem]} "aquele&FEM&SG" =that=
aquelas {[scat det:dem] }"aquele&FEM&PL" =those=
aquele {[scat det:dem] } "aquele&MASC&SG" =that=
aqueles {[scat det:dem] } "aquele&MASC&PL" =those=
essa {[scat det:dem] } "esse&FEM&SG" =this=

det-indef.cut
algum {[scat det:indef]} "algum&MASC&SG"=some=
alguma {[scat det:indef]} "algum&FEM&SG"=some=
algumas {[scat det:indef]} "algum&FEM&PL"=some=
alguns {[scat det:indef]} "algum&MASC&PL"=some=
bastante {[scat det:indef]} "algum&MASC&PL"=quite_some=

det-int.cut
quais {[scat det:int]} "qual&PL" =which_ones=
qual {[scat det:int]} "qual&SG" =which_one=
quanta {[scat det:int]} "quanto&FEM&SG" =how_much=
quantas {[scat det:int]} "quanto&FEM&PL" =how_many=
quanto {[scat det:int]} "quanto&MASC&SG" =how_much=

det-pos.cut
meu {[scat det:pos]}"meu-1S&MASC&SG" =my=
meus {[scat det:pos]} "meu-1S&MASC&PL" =my=
minha {[scat det:pos]} "meu-1S&FEM&SG" =my=
minhas {[scat det:pos]} "meu-1S&FEM&PL" =my=
teu {[scat det:pos]} "teu-2S&MASC&SG" =your=

dialect.cut (expresses dialetais)
nõ {[scat adv]} =no=
boa_tadi =good_afternoon=

ba_tadi =good_afternoon=
gefilte =Yiddish fish=
íguerque =Yiddsh cucumber=

dim.cut
afinadinho {[scat adj]} "afinado-DIM-MASC&SG" =quite_tunned=
alemãzinha {[scat adj]} "alemão-DIM-FEM&SG" =German_child=
baixinho {[scat adj]} "baixo-DIM-MASC&SG" =quite_short=
biza {[scat adj]} "bisavó" =grand_grand_mommy=
bocadinho {[scat adv]} "bocado-DIM" =handful=

fam.cut (expressões da criança, incorporadas pela família)
bií [= disco] {[scat n&DIM][gen fem]} =child's backside=
motiqui [= mosquito] {[scat n][gen masc]} =mosquito=
sanduito [= sanduíche] {[scat n][gen masc]} =sandwich=
Pepé [= Pelé] {[scat n][gen masc]} =Pelé=
tapatinho [= sapatinho] {[scat n]} "sapato-DIM-MASC&SG" =little_shoe=

id.cut (expressões idiomáticas)
alô =hello=
amo(r)_de_Deus
boa_ta(r)de =good_afternoon=
coisa_com_coisa =you're_not_making_any_sense=
coluna_um

interj.cut
ah {[scat interj]} =ah; oh=
aham {[scat interj]} =uh-huh=
ahn {[scat interj]} =huh=
ai {[scat interj]} =ouch=
aiai {[scat interj]} =sigh=

n-abbrev.cut
t_v {[scat n:abbrev]}[gen fem]}} =t_v=

n-comum.cut
abraço {[scat n][gen masc]} =hug=
a(r)co {[scat n][gen masc]} =arch=
área {[scat n][gen fem]} =zone=
adiamento {[scat n][gen masc]} =delay=
adiantamento {[scat n][gen masc]} =precocity=

n-prop.cut
Ana {[scat n][gen fem]}
Andrei {[scat n][gen masc]}
Araguaia {[scat n][gen masc]}
Araguari {[scat n][gen fem]}

Araraquara {[scat n][gen fem]}

num.cut
cinco {[scat num]} =five=
dois {[scat num]} "dois&MASC" =two=
duas {[scat num]} "dois&FEM" =two=
primeira {[scat num]} "primeiro&FEM" =first=
quatro {[scat num]} =four=

onoma.cut
bá, bah, bi, bibi, bum

over.cut (ultrageneralizações efetuadas pela criança)
fazeu@over [= fiz] {[scat v:2]} "faze&PPI&3S" =(he) did=
ponheu@over [=pus] {[scat v:2]} "po&PPI&3S" =(he) put=
punher [= pôr] {[scat v:2]} "po&INF" =to_put=
sabo [= sei] {[scat v:2]} "sabe&PI&1S" =(I) know=

part.cut
aliás =rather=
quer_dizer =rather=
isto_é =rather=
ou_melhor =rather=

phon.cut
abe [: abre@v] {[scat v : 3]} "abri&PPI&3S" =open=
acendi [: acender@v] {[scat v:2]} "acende&INF" =turn_on=
cabô [: acabou@v] {[scat v : 1]} "acaba&PPI&3S" =finished=
Cadil [: Caladril] {[scat n : prop][gen masc]}
cavadô [: gravador] {[scat n][gen masc]} =recorder=

prep-det.cut
à {[scat prep]} "a~det:art|o&FEM&SG" =to the=
às {[scat prep]} "a~det:art|o&FEM&PL" =to the=
ao {[scat prep]} "a~det:art|o&MASC&SG" =to the=
aos {[scat prep]} "a~det:art|o&MASC&PL" =to the=
daquela {[scat prep]} "de~det:dem|aquele&FEM&SG" =of that=

prep-pro.cut
à {[scat prep]} "a~pro:dem|o&FEM&SG" =to the one=
ao {[scat prep]} "a~pro:dem|o&MASC&SG" =to the one=
aos {[scat prep]} "a~pro:dem|o&MASC&PL" =to the ones=
às {[scat prep]} "a~pro:dem|o&FEM&PL" =to the ones=
daquela {[scat prep]} "de~pro:dem|aquele&FEM&SG" =of that=

prep.cut
a {[scat prep]} =to=

abaixo_de {[scat prep]} =under_of=
acerca_de {[scat prep]} =about=
acima_de {[scat prep]} =above=
a_despeito_de {[scat prep]} =in_spite_of=

pro-dem.cut
a {[scat pro:dem]}"o&FEM&SG" =in English it is included in wh form=
aquilo {[scat pro:dem]} "aquilo" =that=
as {[scat pro:dem]} "o&FEM&SPL" =in English it is included in wh form=
aquela {[scat pro:dem]} "aquele&FEM&SG" =that=
aquelas {[scat pro:dem]} "aquele&FEM&PL" =those=

pro-indef.cut
algo {[scat pro:indef]} =something=
alguém {[scat pro:indef]} =someone=
algum {[scat pro:indef]} "algum&MASC&SG"=some=
alguma {[scat pro:indef]} "algum&FEM&SG"=some=
alguma_coisa {[scat pro:indef]} =something=

pro-int.cut
o_que {[scat pro:int]} =what=
que {[scat pro:int]} =what=
quem {[scat pro:int]} =who=
qual {[scat pro:int]} =which_one=

pro-pers.cut
% subject case
eu {[scat pro:pers]} "eu&1S&SUBJ" =I=
tu {[scat pro:pers]}"tu&2S&SUBJ" =you=
% 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=
ele {[scat pro:pers]} "ele&3S&MASC&SG&OBJ" =he;him=
ela {[scat pro:pers]} "ele&3S&FEM&SG&OBJ" =she;her=

pro-pos.cut
meu {[scat pro:pos]} "meu-1S&MASC&SG" =mine=
minha {[scat pro:pos]} "meu-1S&FEM&SG" =mine=
meus {[scat pro:pos]} "meu-1S&MASC&PL" =mine=
minhas {[scat pro:pos]} "meu-1S&FEM&PL" =mine=
teu {[scat pro:pos]} "teu-2S&MASC&SG" =yours=

pro-rel.cut
quem {[scat pro:rel]} =who=
cuja {[scat pro:rel]} "cujo&FEM&SG" =which=
cujas {[scat pro:rel]} "cujo&FEM&PL" =which=
cujo {[scat pro:rel]} "cujo&MASC&SG"=which=

cujos {[scat pro:rel]} "cujo&MASC&PL"=which=

syl.cut (silabação)
Mi^ni^ni^nha {[scat n]}[gen fem&DIM]} "menina&DIM&SG" =part of music title=

tag.cut
vamos {[scat tag]} =c'mon=
viu {[scat tag]} =see=
é {[scat tag]} =yeah=
tá {[scat tag]} =ok=
tá_bom {[scat tag]} =alright=

v.cut
conseguem {[scat v:3]} "consegue&PI&3P" =(they) can=
sentar {[scat v:1]} "senta&INF" =to_sit=
amado {[scat v:1]} "ama&PAR&MASC&S" =loved=(voz passiva)
amado {[scat v:1]} "ama&PAR&N" =loved=(tempo composto c/ter ~ haver)

wp.cut
minha mãe {[scat wp]} =my_mother=
atirei_um_pau_no_gato {[scat wp]} =folk_song=

4.2. Léxico verbal de apoio

As ambiguidades do sistema de verbos do português fazem com que várias possibilidades sejam geradas. Por isso, a fim de limitar a saída do analisador morfológico (Figura 1) à saída correta, desenvolveu-se o procedimento para criação um léxico verbal automático de apoio, para o *corpus* de trabalho – que resolveu várias ambiguidades geradas, sobretudo, pela alomorfia da vogal temática, nas três conjugações, e pela harmonia vocálica que ocorre no radical de verbos da 3.ªC – o qual já foi descrito e discutido (VASILÉVSKI *et al.* 2012). Cabe apenas lembrar que o léxico verbal automático contém somente a forma infinitiva dos verbos, e é usado para a criação da linha mor%. Depois, a partir da linha mor%, é criado o arquivo v.cut, que contém a análise morfológica de todas as formas verbais conjugadas constantes no *corpus*.

Conclusões

Neste artigo, propusemo-nos definir o que é a gramática automática, seguindo-se o histórico da gramática automática do PB. Descrevemos, na continuidade, os procedimentos para a preparação da linha principal para a análise morfológica por regras, examinando os principais desafios encontrados: a falta, às vezes, de constituintes essenciais à frase; os deslocamentos agramaticais; as descontinuidades provocadas pelas pausas plenas ou vazias de processamento; a reduplicação do pronome interrogativo átono "que" com a função de redundância, dada sua fraca saliência perceptual; a superposição de vozes (*overlapping*); as mudanças diacrônicas e o fenômeno de sândi externo. Expusemos, após, os critérios adotados

para alimentar o léxico ao qual são pareados os itens reconhecidos na linha principal, discutindo as soluções para questões difíceis, como a ambiguidade lexical e o resgate da forma infinitiva dos verbos do *corpus* a partir de suas formas conjugadas: todas as classes sintáticas, inclusive inovações específicas relativas à linguagem da criança foram, então, exemplificadas. Com isso, são socializados dois dos três procedimentos para a depreensão da gramática automática do português brasileiro.

ABSTRACT: In this work, we define an automatic grammar, describing the procedures for preparing the main line for rule-based morphological analysis, and exposing the criteria used for feeding the automatic lexicons, to which the items recognized in the main line are matched. The theoretical frameworks are the competitive model (MacWhinney, 2000a, 2008) and Brazilian authors who have studied the current Brazilian Portuguese. The original corpus comes from the doctoral thesis of Scliar-Cabral (1976), and the results already obtained are the formatting of all lines of the corpus, the organization of lexicons and the development of algorithms (Vasilévski, Araújo, 2011, Vasilévski et al., 2012).

Key-words: Automatic Grammar; Brazilian Portuguese; lexicons; ambiguity; CHILDES

Referências

BASÍLIO, M. *Estruturas lexicais do português*. Petrópolis: Vozes, 1980.

BASÍLIO, M. *Teoria lexical*. São Paulo: Ática, 1987.

BASÍLIO, M. *A delimitação de unidades lexicais*. Rio de Janeiro: Grypho, 1999.

BASÍLIO, M. Em torno da palavra como unidade lexical: palavras e composição. *Veredas*, v.4, n.2, 2000.

BASÍLIO, M. *Formação e classes de palavras no português do Brasil*. São Paulo: Contexto, 2004.

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.

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

CASTILHO, A. T. (Org.). *Gramática do português falado*. 3.ed. rev. Campinas: UNICAMP/FAPESP. 2002b, v.III, As abordagens.

CHILDES. *Child Language Data Exchange System*. Disponível em: <http://childes.psy.cmu.edu/>. Acesso em: abr. 2012.

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

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

MacWHINNEY, B. Lexicalist connectionism. *In*: BROEDER, P.; MURRE, J. (Eds.). *Models of language acquisition*. Oxford: Oxford University Press, 2000a, p.9-31.

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

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

MATTOSO CAMARA JR., J. *Problemas de linguística descritiva*. Petrópolis: Vozes, 1971.

MATTOSO CAMARA JR., J. *História e estrutura da língua portuguesa*. 2.ed. Rio de Janeiro: Padrão, 1976.

MATTOSO CAMARA JR., J**.** *Estruturas da língua portuguesa*. 26.ed. Petrópolis: Vozes, 1997 [1970].

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***.** 2.ed.rev. Campinas: UNICAMP/Humanitas, 1999, v. VII, Novos estudos.

NOMENCLATURA GRAMATICAL BRASILEIRA – NGB. (1959) Disponível em <http://www.portaldalinguaportuguesa.org/?action=ngbras>. Acesso em: abr. 2012.

SCLIAR-CABRAL, L. *Introdução à linguística*. Porto Alegre: Globo, 1971.

SCLIAR-CABRAL, L. *A explanação lingüística em gramáticas emergentes*. 1976. Tese (Doutorado) – USP, Pós-graduação em Linguística. São Paulo.

VASILÉVSKI, V.; ARAÚJO, M. J. Tratamento dos sufixos modo-temporais na depreensão automática da morfologia dos verbos do português. *Linguamática*, v.3, n.2, p.107-118, dez. 2011.

VASILÉVSKI, V.; SCLIAR-CABRAL, L.; ARAÚJO, M. J. Automatic Analysis of Portuguese Verb Morphology: solving ambiguities caused by thematic vowel allomorphs. *Proceedings of the 10th International Conference on the Computational Processing of Portuguese (PROPOR)*, Coimbra, Portugal, 2012.

Data de envio: 10/05/2012
Data de aceite: 01/08/2012
Data de publicação: 15/03/2013