Capítulo 6

A anotação gramatical do C-ORAL-BRASIL

Eckhard Bick

1. Introdução

Ao longo dos últimos 20 anos, a Linguística de *Corpus* tem experimentado uma progressiva ampliação do seu foco de pesquisa, para além do estudo da língua escrita e incorporando também estudos da língua oral. Tanto em seu escopo quanto em seu volume essa mudança tem sido bem-sucedida e, hoje, centenas de *corpora* orais estão disponíveis. Entretanto, comumente, essa disponibilidade é restrita a grupos individuais de pesquisa, ou é comprometida pela ausência de mecanismos de busca. Essa segunda limitação está muito ligada a questões oriundas dos processos de anotação de *corpora* – para ser otimamente útil, um *corpus* de fala deve ter um sistema de anotação padronizado e refinado em diferentes níveis, tais como o nível prosódico, a estrutura discursiva etc.; incluindo-se aí a tradicional anotação morfossintática. Neste capítulo, enfocaremos a integração da anotação morfossintática com os níveis prosódicos e discursivos e discutiremos a questão de se e como um etiquetador-*parser*[1] criado primariamente para atender às demandas da escrita pode ser adaptado para o tratamento de dados orais transcritos. O trabalho foi desenvolvido no contexto de pesquisa do projeto de fala espontânea C-ORAL-BRASIL para o português brasileiro (Raso; Mello, 2010), no qual a anotação morfossintática deveria ser adicionada à já existente meta-anotação de ortografia não padrão e ausência de

[1] O termo *parser* será mantido como no texto original em inglês. (N.T.)

pontuação, preservando marcadores do fluxo da fala (p. ex., interrupções, retrações, unidades tonais etc.).

A utilização de anotação automática, isoladamente ou como um passo que antecede à revisão manual, é uma escolha óbvia para um *corpus* deste tamanho (~300.000 palavras). Assim, subprojetos irmãos do C-ORAL-BRASIL, integrantes do C-ORAL europeu,[2] utilizaram etiquetadores estatísticos para essa tarefa, tais como o sistema PiTagger (Moneglia; Panunzi; Picchi, 2004) para a seção italiana, o qual acessava um analisador baseado no léxico, um léxico padrão (106.090 lemas), um *corpus* de treinamento (50.000 palavras) e um pré-dicionário especial cobrindo cerca de 2.000 formas dialetais e não padrão. Para a seção de português europeu, foi utilizado o etiquetador Brill (Brill, 1992), treinado com um *corpus* de português escrito de 250.000 palavras. Apesar de nenhuma anotação sintática de ordem mais alta haver sido tentada para o C-ORAL europeu, outros projetos de *corpus* de fala já optaram por anotações de tipo *treebank*[3] completas, tais como o *treebank* árabe, descrito por Maamouri *et al.* (2010), que combinou seleção manual das sugestões do analisador, seguida do estágio de análise sintática automática. Porém, o *treebank* árabe foi construído com língua midiática, e não a partir de entrevistas ou diálogo espontâneo; assim, uma comparação direta não pode ser feita com o C-ORAL-BRASIL, já que aquele necessitou de muito menos meta-anotação discursiva e de formas não padrão em sua transcrição de dados telejornalísticos.

No nosso próprio trabalho, utilizamos o *parser* PALAVRAS (Bick, 2000) como ponto de partida. O PALAVRAS é um *parser* baseado na Gramática Constritiva[4] (CG) utilizado majoritariamente para a anotação de dados escritos, mas tem demonstrado grande robustez em sua aplicação a diferentes gêneros (como, por exemplo, no projeto Linguateca[5] e nos *corpora*[6] CorpusEye). Com adaptações lexicais e vários programas de filtros, o *parser* já foi utilizado para variedades não padrão das línguas, tais como textos históricos (Bick; Módolo, 2005). O paradigma da CG (Karlsson *et al.*, 1995),

---

[2] Projeto C-ORAL-ROM: <http://lablita.dit.unifi.it/coralrom/>.

[3] Em português europeu, os *treebanks* são chamados de Floresta Sintáctica (cf. Linguateca: <www.linguateca.pt>). (N.T.)

[4] *Constraint Grammar* no original em inglês. Será mantida a abreviatura CG ao longo do capítulo para se referir à mesma. (N.T.)

[5] <www.linguateca.pt>.

[6] <www.corp.hum.sdu.dk>.

ao qual o PALAVRAS está associado, pode ser descrito como um dualismo entre uma metodologia modular robusta para o Processamento da Linguagem Natural (PLN), por um lado, e uma convenção linguístico-descritiva, por outro, codificando análises linguísticas como etiquetas baseadas em ocorrências e estruturas de dependência mediadas por sua função. Tanto o método quanto a tradição descritiva oferecem um número de vantagens formais para a anotação de dados linguísticos não padrão, como os encontrados na fala espontânea. Primeiramente, porque os sistemas da CG têm uma arquitetura modular, com uma clara separação de léxicos, analisadores e gramáticas (conjuntos de regras) para níveis sucessivos de análise; é relativamente fácil acrescentarem-se léxicos especializados ou filtros morfológicos, assim como adicionar módulos específicos de gramática. Em segundo lugar, a anotação da CG baseada em ocorrências, na qual a mesma informação estrutural de alto nível é estritamente baseada em ocorrências, permite a um projeto de *corpus* manter vários níveis de anotação em paralelo (tais como marcadores discursivos em oposição a fronteiras de orações), possibilitando até mesmo que regras que lidam com um nível façam referência a etiquetas de outro nível. Vários projetos de anotação de fala fazem uso dessas vantagens, tais como Müürisep e Uibo (2006) para o estoniano e Lindstad *et al.* (2009) para o Corpus do Dialeto Nórdico, apesar de este último utilizar uma técnica híbrida, na qual uma CG para textos escritos foi usada para anotar uma porção dos dados de fala da área de Oslo, que foi então manualmente corrigida e utilizada para treinar um Etiquetador de Decisões Estruturais-Arbóreas (Schmid, 1994) para uso com outros dialetos noruegueses. No contexto do C-ORAL-ROM, o *subcorpus* espanhol usou regras baseadas na CG para a desambiguação das partes do discurso na saída morfológica do sistema GRAMPAL (Moreno; Guirao, 2003); e para o próprio *parser* PALAVRAS, Bick (1998) relatou experimentos iniciais com soluções estritamente baseadas na CG em relação à anotação morfossintática do *corpus* brasileiro NURC.[7]

2. O ponto de partida: o *parser* PALAVRAS

Tecnicamente o *parser* PALAVRAS é uma cadeia de conjuntos de regras da CG, sucessivamente lidando com níveis cada vez mais altos (profundos) de análise, progredindo da desambiguação morfológica e etiquetagem das

[7] Norma Linguística Urbana Culta (Castilho, 1993).

partes do discurso (PoS), para o mapeamento de funções sintáticas e relações de dependência, para a anotação de papéis semânticos, Reconhecimento de Entidades Nomeadas e módulos de aplicação orientada. O insumo para essa cadeia de gramáticas é fornecido por um pré-processador/atomizador[8] e um programa de análise morfológica sustentado por amplos léxicos cobrindo paradigmas flexionais, no estilo da CG, como etiquetas ligadas a ocorrências nas linhas de leitura. Linhas de leitura ambíguas para uma dada palavra são chamadas de um *coorte*, como seria o caso para as típicas ambiguidades verbo-nominais portuguesas envolvendo as terminações -a e -o:

"<casa>"
"casa" <build> N F S ('residência')
"casar" <vt> <vi> <com^vp> <vr> <com^vrp> <vH> V IMP 2S VFIN ('casa!')
"casar" <vt> <vi> <com^vp> <vr> <com^vrp> <vH> V PR 3S IND VFIN ('ele/ela casa')

"<accordo ALT acordo>"
"acordo" <sem-c> <+com> <+sobre> <+entre> <de+> <+n> ('contrato')
"acordar" <ve> <vt> <vK> V PR 1S IND VFIN ('eu acordo')

Uma distinção é feita entre etiquetas primárias, as quais são eleitas para desambiguação (ex. N, V), e etiquetas secundárias que não são destinadas (ou pelo menos não neste nível) à desambiguação (etiquetas <...>), que servem como pistas contextuais para regras da CG no processo de desambiguação primária. Assim, uma etiqueta de verbo transitivo <vt> e uma etiqueta de nome humano <H> podem auxiliar na designação de função de sujeito ou objeto a um nome na sentença.

A saída anotada do PALAVRAS oferecerá preferencialmente uma única etiqueta primária de cada tipo[9] (apesar de as sequências de etiquetas

---

[8] *Tokenizer*. (N.T.)

[9] Para orações dependentes, duas notações convencionais podem ser escolhidas. Enquanto o novo padrão de codificação, compatível com o VISL, codifica a função de dependente com uma única etiqueta no verbo núcleo (primeiro verbo) da dependente, a convenção do PALAVRAS original usa duas etiquetas para dependentes, ou no subordinador (dependentes finitas) ou no verbo (dependentes não finitas), uma etiqueta é interna (@) e a outra externa (@#) indicando a função da dependente como um todo.

poderem ser complexas, como no caso de uma sequência de etiquetas de número e gênero):

O <artd>	DET M S	@>N	#1->3
último	ADJ M S	@>N	#2->3
diagnóstico	N M S	@SUBJ>	#3->9
elaborado	V PCP2 M S	@IMV @#ICL-N<	#4->3
por	PRP	@<PASS	#5->4
a <artd>	DET F S	@>N	#6->7
Comissão=Nacional	PROP F S	@P<	#7->5
não	ADV	@ADVL>	#8->9
deixa	V PR 3S	@FMV	#9->0
dúvidas	N F P	@<ACC	#10->9
$.			#11->0

Os exemplos mostram uma anotação no nível da estrutura de árvore, com campos nas etiquetas relativos à parte do discurso (N=nome, ADJ=adjetivo, V=verbo etc.), morfologia (M=masculino, F=feminino, S=singular, P=plural etc.) e função sintática (@SUBJ=sujeito, @>N=pré-nominal, @#ICL-N< = relativa, oração pós-nominal não finita, @<ACC =objeto acusativo, @P< argumento de preposição, @FMV=verbo finito principal, @IMV=verbo não finito principal). As ligações de dependência já estão implícitas no nível da função sintática, através dos marcadores de direção > e <, apontando em direção ao núcleo do sintagma ou oração. A árvore de dependência completa é mostrada através de "etiquetas de arco"[10] #n->m, nas quais n é a ID da ocorrência, e m a ID da sua mãe de dependência. Estruturas sintáticas em árvore, como essas, podem ser refeitas em diferentes formatos. Para árvores completas, o PALAVRAS oferece, por exemplo, os colchetes de constituintes chomskianos tradicionais, anotação do PENN *treebank*, XML do TIGER *treebank* e marcação de dependência XML do MALT.

[10] Aspas no original: *arc tags*. (N.T.)

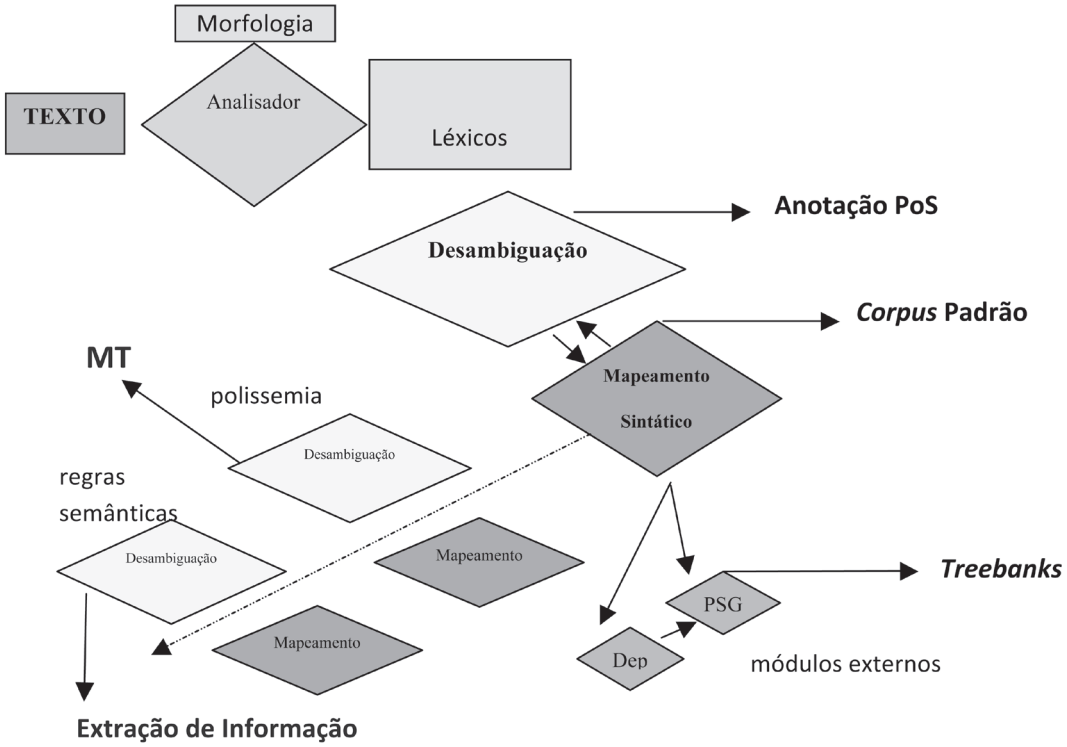

Figura 1 - Gráfico de fluxo do processador

O PALAVRAS utiliza cerca de 6.000 regras contextuais da CG que removem, selecionam, adicionam, mapeiam ou substituem etiquetas/leituras. Fora outras etiquetas (associadas a qualquer outra palavra na sentença), o formalismo da CG permite que as regras façam referência à informação estatística/numérica relacionada a palavras ou leituras, combinem expressões regulares com formas de palavras, etiquetas e lexemas, ou unifiquem traços categoriais ao longo dos constituintes. Até mesmo regras de reescrita gerativistas, apesar de serem de alguma forma estranhas à metodologia reducionista do formalismo da CG, podem ser expressas com o auxílio dos assim chamados *templates*.

A maior parte das regras, entretanto, é de desambiguação, como aquelas a seguir, que removem a leitura de verbo finito (VFIN) se há uma leitura de preposição segura (C) à esquerda:

REMOVE VFIN IF (-1C PRP); # remove a[11]

Apesar de esse tipo de contexto poder ser capturado por modelagem de n-gramas de Modelos de Markov probabilísticos não observáveis (HMM),[12] muitas das regras da CG têm escopo global sobre a sentença e são consideravelmente mais potentes. Considere, por exemplo, a seguinte regra de unicidade, que diz que uma palavra não pode ser um verbo finito se já existe um verbo finito em alguma outra posição (*) à esquerda (*-1), sem uma fronteira de oração (CLB) ou coordenador (KC) no meio (BARRIER):

REMOVE VFIN (*-1 VFIN BARRIER CLB OR KC)

Dadas a arquitetura e a metodologia das regras do *parser*, três desafios podem ser identificados no que diz respeito à aplicação aos dados da fala, afetando a revocação lexical por um lado (2) e a desambiguação contextual, por outro (1,3). Os problemas encontrados se assemelham de várias maneiras àqueles presentes na anotação de dados linguísticos históricos (Bick; Módolo, 2005):

1. Como manter a meta-anotação do *corpus* separada das camadas de anotação não gramatical, porém ainda assim oferecer insumo de texto corrente para o *parser* e seu analisador?
2. Como adaptar o léxico e/ou alterar as formas das palavras para permitir que o insumo seja reconhecido como português brasileiro padrão, contemporâneo e escrito, enquanto, ao mesmo tempo, mantém as formas de transcrição oral?
3. Como inserir quebras sintáticas na ausência de pontuação, a fim de se permitir que sejam definidas janelas para desambiguação contextual?

Trataremos dessas questões nas seções 3, 4 e 5, respectivamente.

---

[11] As regras do *parser* serão mantidas em sua forma original em inglês. (N.T.)

[12] *Hidden Markov Models*. (N.T.)

3. A normalização do fluxo textual

O C-ORAL-BRASIL utiliza símbolos e convenções de codificação para lidar com questões do fluxo de dados, tais como tomadas de turno, quebras prosódicas, sobreposição de falas, retrações e interrupções. Tais codificações são feitas em formatos não alfanuméricos (<, /, +), ou estão fora do enunciado (nomes dos falantes); dessa forma, elas ou não podem ou não devem ser analisadas pelo *parser*. Para ao mesmo tempo manter essa metainformação e oferecer somente insumo textual ao *parser*, optamos por uma anotação de dois níveis, na qual a metainformação é "armazenada" em colchetes angulares em linhas separadas como metamarcação do *corpus*, semelhante a, por exemplo, marcadores <source>, <s> e <p> em anotação de textos escritos. A anotação do PALAVRAS é transparente a esse tipo de marcação e não mudará, removerá ou tentará analisá-la. Considere o seguinte exemplo de dois turnos, primeiramente na anotação original do C-ORAL-BRASIL e depois no formato vertical da CG, depois do processamento.

***LEO:** o Juninho **<foi>** / /
***GIL:** **<ô / mas>** / voltando à questão / **falando em** [**/2**] e também falando em povo mascarado / esse povo do Galáticos é muito palha / eu acho que es nũ deviam mais participar / e **<tal>** / /

<LEO:>

o [o] <artd> DET M S @>N

Juninho [Juninho] <hum> <newlex> <*> PROP M S @ SUBJ>

<overlap-start>

foi [ser] <fmc> V PS 3S IND VFIN @FMV

<overlap-stop>

$;

<GIL:>

<overlap-start>

ô [ô] <newlex> IN @ADVL

$,

mas [mas] KC

<overlap-stop>

$, voltando [voltar] V GER @IMV @#ICL-ADVL>

a [a] <sam-> PRP @<PIV

a [o] <-sam> <artd> DET F S @>N

questão [questão] <ac> N F S @P<

$,

<retract: falando_em>

e [e] KC

```
também  [também] ADV @
    ADVL>
falando  [falar] <vH> V GER @
    IMV @#ICL-<ADVL
em  [em] PRP @<PIV
povo  [povo] <HH> N M S @P<
mascarado [mascarar] <vH>
    V PCP M S @N<
$,
esse  [esse] <dem>
    DET M S @>N
povo  [povo] <HH>
    N M S @SUBJ>
de  [de] <sam-> PRP @N<
o [o] <-sam> <artd>
    DET M S @>N
Galáticos [Galáticos] <org>
    <newlex> <*>
    PROP M P @P<
é  [ser] <vK> <fmc>
    V PR 3S IND VFIN @FMV
muito [muito] <quant>
    ADV @<ADVL
palha  [palha] <cm>
    N F S @<SC
$,
eu  [eu] PERS M/
    F 1S NOM @SUBJ>
acho  [achar] <vH> <fmc>
    V PR 1S IND VFIN @FMV
que  [que] KS @SUB @#FS-<ACC
es OALT eles  [eles]
    PERS M 3P NOM @SUBJ>
nũ OALT não  [não] ADV @<
    ADVL
deviam  [dever] V IMPF 3P IND
    VFIN @FAUX
mais  [mais] ADV @<ADVL
participar  [participar] <vH>
    V INF @IMV @#ICL-AUX<
$,
e  [e] KC
<overlap-start>
tal  [tal] <diff> <KOMP>
    DET M/F S @<OC
<overlap-stop>
$;
```

Aqui, somente as linhas não iniciadas com '<' são parte da anotação morfossintática. Nomes de falantes são metaetiquetas separadas <GIL:>, e sobreposições (<...>) são marcadas com <overlap-start> e <overlap-stop>. É claramente vantajoso para o *parser* que as retrações sejam pré-marcadas manualmente em colchetes no seu ponto inicial, sinalizando assim o número de palavras retraídas. O nosso módulo pré-processador somente necessita de eliminar as palavras em questão do nível de superfície para permitir um processamento sintático mais fácil. Repetições de palavras ou autocorreções, se permitidas a permanecerem no nível de superfície, seriam problemáticas para as regras da CG em todos os níveis, interferindo não

somente com a implementação de universais linguísticos como o princípio de unicidade, mas também com a adjacência de classes de palavras e regras de concordância.

Como pode ser visto no exemplo, as palavras deletadas da superfície serão armazenadas em uma etiqueta especial <retract:...>, mantendo o princípio da anotação em dois níveis, através do qual dois níveis de anotação são separados, mas não mutuamente exclusivos. O mesmo procedimento é utilizado para as chamadas não palavras, que podem ser de dois tipos: primeiramente algumas sequências de não palavras de superfície sem marcação especial ('hhh' e 'xxx'), e, em segundo lugar, palavras incompletas (contrações), que são marcadas com um sinal inicial &.

*GIL: **hhh** eu tenho **&dire**
<GIL:>
<nonword:hhh>
eu [eu] PERS M/F 1S NOM @SUBJ>
tenho [ter] <fmc> V PR 1S IND VFIN @FMV
<nonword:&dire>

Uma vez que a reescrita de marcações de superfície acima como meta-etiquetas do *corpus* necessita acessar texto corrente e não palavras individuais, e é designada para se sobrepor à própria atomização[13] do PALAVRAS, ela deve ser feita por um pré-processador, anteriormente ao processo de atomização. Em termos de sucessão de módulos, este também é o ponto no qual o tratamento de multipalavras do PALAVRAS pode ser sobreposto. Assim, expressões multipalavras (MWEs) de arquivos específicos do *corpus* são utilizadas para criar multipalavras tais como '*emepê=três*' (MP3), '*al=dente*', '*air=bags*', evitando que o PALAVRAS efetue análises parciais, e ao contrário, designe o léxico especificamente etiquetado do *corpus* com formas apropriadas. Assim, '*air=bags*' não receberá uma análise original do sistema do PALAVRAS (*default*, nome singular), mas sim uma leitura de Nome Plural proveniente do léxico do nosso *corpus* (N M P).

Uma complicação adicional surgiu do fato de marcações de sobreposição e retração poderem estar aninhadas e/ou sobrepostas, como mostra o exemplo a seguir, o que requer um ordenamento cuidadoso de pareamentos

[13] O termo *tokenization* do original foi aqui traduzido como atomização, em consonância com o termo utilizado pelo portal LINGUATECA (<www.linguateca.pt>). (N.T.)

em cadeia, por exemplo, para evitar que retrações se tornem "invisíveis" dentro de marcadores destextualizados para falantes em sobreposição. Adicionalmente, uma vez que sobreposições e não palavras podem aparecer contidas no escopo de uma retração, elas mudariam a contagem do número de palavras dessa última se removidas cedo demais no processo, e possivelmente afetariam palavras legítimas mais à esquerda.

```
*GIL: <eu &a [/2] eu acho que é> esse [/2] é esse aqui o' // <&he> +
<GIL:>
<overlap-start>
<retract:eu_&a>
eu      [eu] PERS M/F 1S NOM @SUBJ>
acho [achar] <vH> V PR 1S IND VFIN @FMV
que     [que] KS @SUB @#FS-<ACC
<retract:é>_esse>
é       [ser] <vK> V PR 3S IND VFIN @FMV
esse    [esse] <dem> DET M S @<SC
aqui    [aqui] ADV @N<
o' OALT olha [olhar] <vH> V PR 3S IND VFIN @FMV
$;
<overlap-start>
<nonword:&he>
<overlap-stop>
$
```

Deve-se notar que a sobreposição de colchetes não inclusivos do tipo <a> representa um problema geral de anotação, até mesmo para esquemas de codificação xml elaborados, uma vez que os últimos não preveem estruturas arbóreas não projetivas (sobrepostas); assim, a notação da CG escolhida aqui pode ser rotulada como uma solução bastante robusta.

A língua portuguesa, especialmente em sua modalidade falada, utiliza uma construção de foco especial com *ser ... que*, que em alguns casos está formalmente sujeita à análise com uma oração relativa absoluta (*o que*), porém, na maioria dos casos, o uso determina uma análise mais simples e funcional *é=que*, *foi=que*, ou simplesmente *é* como uma partícula de foco, inserida antes do constituinte focalizado:

***é** uma cerveja **que** quero*
→ *uma cerveja **é que** quero*
→ *quero **é** uma cerveja*

No *corpus* C-ORAL-BRASIL, a partícula de foco é=*que* ocorre 380 vezes, em ~2% dos turnos, mas é transcrita como *que*; já que o PALAVRAS leria um *que* ordinário como uma conjunção ou um relativo, as regras teriam dificuldades devido à ausência de uma oração subordinada. Assim, o pré-processador de normalização tenta parear sequências de palavras *qu* e *que* (*que que*, *quando que*, *quanto que*, *quem que*, *onde que*), e insere o *é=que* padrão, ao mesmo tempo que retém a sequência substituída em uma metaetiqueta <elision: ...>:

*LEO: *<beleza> // <então a gente já sabe em quem que a gente vai colocar a> culpa //*

<LEO:>

<overlap-start>

beleza [beleza] <am> N F S @NPHR

<overlap-stop>

$;

<overlap-start>

então [então] <kc> ADV @ADVL>

a [o] <artd> DET F S @>N

gente [gente] <HH> N F S @SUBJ>

já [já] ADV @ADVL>

sabe [saber] <fmc>
V PR 3S IND VFIN @FMV

em [em] PRP @ADVO>

<ellision:quem_que>

quem [quem] <interr> SPEC M/F S/P @P<

é=que [é=que] <foc> ADV @<FOC

a [o] <artd> DET F S @>N

gente [gente] <HH> N F S @SUBJ>

vai [ir] V PR 3S IND VFIN @FAUX

colocar [colocar] <vH> V INF @IMV @#ICL-AUX<

a [o] <artd> DET F S @>N

<overlap-stop>

culpa [culpa] <am> N F S @<ACC

$;

Uma vez que o PALAVRAS não é simplesmente um etiquetador morfossintático, mas é também um *parser* que fornece análise arbórea profunda, ele necessita que preposições e pronomes preencham suas respectivas lacunas sintáticas (em SPs e SNs), mesmo no caso de formas superficiais contraídas como *deles* (*de eles*), *num* (*em um*), *pelos* (*por os*). A resolução de contrações torna a estrutura sintática mais transparente e faz com que as generalizações linguísticas sejam mais simples. Assim, ela traz vantagens tanto para a desambiguação da gramática (a valência verbal pode "ver" uma dada preposição, os nomes podem "ver" os seus artigos) quanto para a facilitação de tarefas como a extração de SNs, o alinhamento translinguístico de palavras e a extração de padrões colocados em lexicografia. Porém, se por um lado o *parser* automaticamente separa contrações com as preposições *de* (~52%), *em* (~37%), *a* (2,6%) e *por* (1,7%), assim como formas históricas com *com* (2%) e algumas contrações frequentes de *para* (5%), ele não cobre todas as combinações desta última forma e, obviamente, perde as contrações cuja segunda parte é feita de formas não padrão, como *naquea* (*em aquela*). Assim, as formas remanescentes tiveram que ser expandidas pelo pré-processamento do C-ORAL-BRASIL, ou (a) antes ou (b) depois do processo de atomização do PALAVRAS. Para a pré-atomização, um pareamento via expressão regular foi usado, e uma metaetiqueta <contraction:...> foi inserida na frente da expansão; todas foram expansões de duas partes, listadas abaixo com sua frequência no *corpus*:

pa (477), pro (122), co (23), pros (20), prum (18), pos (18), ca (15), pras (14), cum (9), cos (7), des (8), cos (7), pum (6), puma (5), naquea (5), cuma (5), pruma (4), dea (3), daquea (3), pas (1), naques (1), daqueas (1).

eu falei isso naquea reunião lá

	em [em] PRP @<ADVL
eu [eu] PERS M/F 1S NOM @SUBJ>	
falei [falar] <vH> V PS 1S IND VFIN @ FMV	aquela [aquele] <dem> DET F S @>N
isso [isso] <dem> SPEC M S @<ACC	reunião [reunião] <occ> N F S @P<
<contraction:naquea>	lá [lá] ADV @N<

O caso mais problemático foi *pra*, porque essa forma é ambígua: ela pode ser tanto a versão contraída de *para*, quanto uma contração de duas palavras – *para a* –, o que requer desambiguação contextual.

/ só **pra** eles mesmos / / *(=para)*
/ / **pra** próxima taça / *(=para a)*

A pós-atomização (programa *coral.inter*) foi utilizada para as contrações que são menos regulares e/ou mais difíceis de serem pareadas com expressões regulares. Esses casos foram retirados do léxico normalizado do C-ORAL-BRASIL, e suas partes foram numeradas como formas de palavras e marcadas com a etiqueta normalizada OALT (cf. seção 4); assim, em princípio permitiriam qualquer número de partes:

*pa despesa é bastante / **né** / /*

pa OALT	pra [para] <sam-> PRP @ADVL>	$,
a	[a] <artd> <-sam> DET F S @>N	<slash>
despesa	[despesa] <mon> N F S @P<	**né OALT não** [não] ADV @ ADVL>
é	[ser] <vK> V PR 3S IND VFIN @FMV	**né-2 OALT é** [ser] V PR 3S IND VFIN @FMV
bastante	[bastante] <nh> ADJ M/F S @<SC	$;

Observe-se que a padronização de *pa* é *pra*, e não *para*, deixando-se assim que o PALAVRAS resolva a ambiguidade *para=a/para*.

4. A normalização do léxico e da ortografia

Para designar uma etiqueta morfológica e uma hipótese sobre a classe de palavra, o PALAVRAS tenta reconhecer palavras desconhecidas como (1) derivações de afixos ou (2) variações de formas padrão, ou uma combinação dessas duas possibilidades. Mesmo para dados oriundos da língua escrita, (2) é um fator de robustez importante por causa das diferenças ortográficas entre o português brasileiro e o europeu, como em (a) e (b), variação oi-ou etc.; assim como para reconhecer textos com uma ortografia considerada

obsoleta por causa de reformas ortográficas. Até mesmo alguns erros tipográficos e formas históricas podem ser reconhecidos dessa maneira. A forma padrão reconhecida será justaposta à forma original através de uma etiqueta ALT:

(a) dicção ALT dição [dição] <sem-s> N F S
(b) négócio ALT negócio [negócio] <act-d> N M S

Assim, o esquema completo com a tipologia de etiquetagem para o C-ORAL-BRASIL será o seguinte:

wordform (ALT normalization) [lemma] <secondary tags> PoS MORPHOLOGY @SYNTAX

Em alguns casos raros, o analisador trocará uma palavra desconhecida, porém existente, de forma a bater com uma análise derivacional, por exemplo, ele lerá *hemácia* como *hemacia* (*hem-ac-ia*). Apesar de esse método conduzir a um erro de lexema, na maioria dos casos ele designará uma análise de PoS e morfologia corretas.

Para o projeto C-ORAL-BRASIL, porém, a padronização normal foi avaliada como não sendo suficiente, primeiramente porque algumas formas de palavras orais foram transcritas foneticamente como ocorriam,[14] criando algumas vezes diferenças irrecuperáveis a partir de uma ortografia padrão, ou o risco de ambiguidade. Como uma consideração colateral, também queríamos dar conta de lacunas lexicais devidas a formas dialetais ou a formas raras. Consequentemente, dois novos módulos foram adaptados à cadeia de programação do PALAVRAS, ambos com um arquivo de léxico mantido manualmente como insumo. O primeiro programa (*coral.inter*) lida com padronizações específicas ou sistemáticas e é rodado depois do pré-processamento, mas antes da análise morfológica; o segundo programa

[14] A transcrição dos dados de fala tem que encontrar um equilíbrio entre a padronização e a fidelidade fonética. Se a padronização for muito reduzida, isso tornará o *corpus* difícil de usar e de se fazerem buscas, análises de frequência lexical ou estudos sobre a ordem das palavras. Muito pouca fidelidade fonética, por outro lado, removerá alguns dos traços e padrões pelos quais poderíamos estar interessados no *corpus*. Assim, perguntas como o quão comum é *im* usado como diminutivo ou quão comum é a queda do *s* final na flexão verbal obviamente não podem ser respondidas se a normalização integral for utilizada. Assim, somente uma anotação de dois níveis, como a que propomos, permitindo tanto buscas por forma e por categoria ao mesmo tempo, pode ter esperança em obter o melhor das duas possibilidades transcritórias.

(*postlex_pt*) é um analisador morfológico comum, com seu próprio léxico e regras de flexão, sobrepondo-se à análise do PALAVRAS, removendo o risco de erro criado por leituras heurísticas.

Um exemplo de normalização sistemática é a adição de *-s* à primeira pessoa do plural dos verbos (*comemoramo -> comemoramos*, *encontramo -> encontramos*), o que é feito pelo *coral.inter* através de cadeias de pareamento e de um léxico de formas completas que ajuda a evitar falsas adições de *-s* a, por exemplo, nomes como *bálsamo, dínamo, esparramo*. A variação *l-r* (*glandão - grandão*) também foi contemplada, mas mostrou ser negligenciável em termos quantitativos.

Cerca de 700 normalizações foram listadas em um arquivo de léxico especial,[15] e, apesar de o analisador padrão ter condições de tratar uma certa proporção dele com seus próprios recursos para classes de palavras, o tratamento lexical também nos permitiu adicionar formas base corretas e até mesmo classificação semântica. Um exemplo muito fonético são as abreviaturas (a1-3), em que mesmo formas de plural (a2) e pronúncias não padrão (a3) foram consideradas.[16] Outros grupos incluem flexão não padrão (d1-3) e derivação (c1-2). Finalmente, mudanças em início de palavras, como aférese (b2-4), tinham que ser incluídas para evitar que fossem consideradas como (provavelmente) nomes singulares.

(a1)	emedebê	MDB
(a2)	emeeles	ml
(a3)	emitivi	MTV
(b1)	envinha	vinha
(b2)	garrou	agarrou
(b3)	inda	ainda
(b4)	roz	arroz
(c1)	espim	espinhos
(c2)	ladim	ladinho

[15] A maior parte do conteúdo original tanto do arquivo de normalização quanto do arquivo de adição ao léxico foi providenciada por um dos autores do C-ORAL-BRASIL, Heliana Mello, seguida de checagem de consistência e compatibilidade para os itens individuais, garantindo cobertura completa de etiquetas e prevenindo interferências indesejáveis com o léxico principal do PALAVRAS.

[16] O PALAVRAS lida com escrita fonética também, mas somente para formas de base do tipo (a1), e através da análise de letras como "sufixos" (<DERS>): emedebê "**M**" <DERS **D**> <DERS **B**> b.

(d1) estudemo estudamos
(d2) fazido feito
(d3) fize fiz

As formas originais foram mantidas, porém as formas padronizadas foram inseridas com um prefixo OALT:... e é à forma padrão que as etiquetas de anotação se referem:

meninim OALT menininho [menino] <DERS> N M S

O léxico padronizado também inclui sequências de multipalavras (*a'=aqui -> olha=aqui*, *c'=ocês -> com=vocês*); esse é o motivo pelo qual o pré-processador atomizador precisa ter acesso ao arquivo. Uma vantagem de se ter a normalização de mulipalavras é que as partes individuais oferecem contexto de desambiguação umas às outras, permitindo, por exemplo, o reconhecimento de *a'* como *olha'*, ao invés de uma leitura preposicional ou de determinante, assim como permite a resolução de *n'* como *não* ou *em* nas formas *n'=era* e *n'=ocê*, respectivamente.

O segundo programa adicional, o analisador de sobreposição, é considerado mais sofisticado que o programa de normalização e permite tanto as entradas de formas completas quanto as de formas básicas em seu léxico (*newlex_pt*). Flexões regulares de nomes, adjetivos e verbos serão reconhecidas a partir da própria forma base, mas todas as formas irregulares têm que ser entradas separadamente. Como para o léxico padronizado, as entradas de multipalavras estarão também visíveis ao pré-processador para atomização (d1, b).

No próprio léxico (atualmente, cerca de 2.000 entradas), devido à boa cobertura do PALAVRAS, há muito poucos nomes regulares do português, e aqueles que há poderiam em sua maioria ser reconhecidos pela análise derivacional do PALAVRAS (a1). Mesmo assim, para algumas formas complexas flexionadas (a2-3), pode ser útil evitar a possibilidade de competição de uma análise heurística, por exemplo, *caça-talentos* como uma forma pluralizada *versus* singular-flexionada. Ademais, o *corpus* contém um grupo de palavras estrangeiras que provavelmente são nomes no singular, mas podem ter terminações que desencadeiam uma análise heurística para o português, como algo do tipo, por exemplo, de *remote* (c1). Ainda mais importante é listar palavras estrangeiras que não sejam nomes, como verbos (c3), adjetivos (c4) ou advérbios (c5), mas essas entradas levantam

dois problemas que teriam que ser resolvidos se o léxico fosse usado em um cenário mais geral (isto é, para outros *corpora*). Primeiramente, palavras estrangeiras teriam que ser especificadas para *todas* as suas leituras, e não somente para aquela que ocorre no *corpus*, por exemplo, *shift* (c4), tanto como nome como verbo. Em segundo lugar, também as entradas estrangeiras precisariam de uma morfologia completa, se fosse para elas interagirem completamente com suas regras e seu contexto portugueses (por exemplo, questões ligadas à concordância). Esta última observação já foi tomada em consideração através da adição semiautomática dos traços masculino e singular (N M S) às entradas nominais sem pré-entrada morfológica; porém uma estratégia semelhante seria mais difícil para verbos e adjetivos dado o fato de a língua inglesa subespecificar o número para adjetivos e, em muitas formas, a finitude verbal.

A maior parte do léxico, entretanto, cerca de dois terços de todas as entradas, são nomes próprios (e1-3). Apesar de estes poderem ser seguramente reconhecidos pelo PALAVRAS, o seu gênero (e possivelmente, número) não é fácil de ser adivinhado (ex. *TIM* como feminino), e a adição de uma leitura de protótipo semântico (ex. <hum>=human, <org>=organization, <Lciv>=town or state) forneceu um contexto semântico valioso para as regras da CG, permitindo, por exemplo, a unificação do traço ±HUM nos verbos em seus sujeitos, condutivas à desambiguação de classes de palavras ou função sintática.

(a1) fazeção <activity> N F S
(a2) zenes N M P # termo de jogo
(a3) caça-talentos N M S
(a4) superbonitinha ADJ F S
(a5) superbem-arrumada ADJ F S
(b) mil-oitocentos-e=vovó=gostosa NUM M/F P
(c1) remote N M S # estrangeirismo
(c2) completed ADJ M/F S/P # estrangeirismo
(c3) save V # estrangeirismo
(c4) shift N M S # estrangeirismo
(c5) anche ADV # estrangeirismo
(d1) tu=tu X # onomatopeia
(d2) tuf X # onomatopeia
(e1) Titina <hum> PROP F S

(e2) TIM <org> PROP F S # operadora de telefonia
(e3) Timoftol <cm-rem> PROP M S
(f) agadê N M S # HD (harddisk)

Em princípio, o módulo de sobreposição do léxico poderia ser considerado como uma extensão geral ao léxico para o PALAVRAS, uma vez que a maioria das adaptações específicas, como as previamente mencionadas abreviações foneticamente grafadas (f), ou a forma feminina dos adjetivos (a4-5), apesar de não estarem em formato padrão, também não perturbam o sistema morfológico do PALAVRAS. Por outro lado, a lista contém algumas poucas entradas sem uma classe de palavras regular, como as onomatopeias *tu tu* e *tuf* (d1-2), e o tratamento de expressões numéricas como um todo (b), as quais estão em conflito com a abordagem mais analítica usada pelo PALAVRAS. Dessa forma, esses tipos de palavras, assim como o potencial de ambiguidade das palavras estrangeiras, devem ser checados em relação à sua consistência antes de serem integrados ao léxico de outros *corpora*.

Originalmente, a extensão lexical do C-ORAL-BRASIL era intencionada a ser uma "substituição dura", isto é, a ideia havia sido a de usar a análise oferecida pela extensão *ao invés* da análise original do PALAVRAS, assumindo-se que esta última seria heuristicamente incorreta ou subespecificada. Porém, a introdução de uma nova leitura, inspirada pela inspeção do *corpus*, porque o PALAVRAS não forneceu a análise desejada em um enunciado em particular, não significa que, necessariamente, para palavras ambíguas, o PALAVRAS forneceria uma análise incorreta em todas as instâncias (isto é, em outros contextos). E, uma vez que é mais difícil para um humano sugerir um coorte completo de leituras ambíguas do que para um computador (um cérebro simplesmente filtra alternativas contextuais sem sentido), uma solução mais conservadora teve que ser tomada, por isso o léxico do C-ORAL-BRASIL *adiciona* às sugestões do próprio PALAVRAS, ao invés de totalmente substituí-las. Uma vez que isso é feito *antes* da desambiguação da CG, as regras gramaticais, então, têm a chance de escolher a melhor leitura contextualmente. Ao mesmo tempo, foi introduzido um viés que permitiu ao léxico do C-ORAL-BRASIL (marcado como <newlex>) substituir mais facilmente as leituras do PALAVRAS marcadas como heurísticas[17] do que aquelas apoiadas no léxico nuclear do PALAVRAS e

[17] Ex.: Com etiquetas <heur> ou etiquetas derivacionais (<DERS>, <DERP>), ou sem uma etiqueta de frequência.

em regras flexionais. Um exemplo desse tipo de interferência ambígua não intencional é a palavra *pô*, listada no léxico C-ORAL-BRASIL como uma interjeição. Devido ao fato de as convenções gerais de transcrição do C-ORAL-BRASIL permitirem a flexão de plural em interjeições, isso significa que a forma *pôs* também seria etiquetada como uma interjeição, competindo com o seu sentido verbal comum.[18] A sentença-exemplo mostra, de modo a rastrear a regra, como o módulo morfológico da CG desambigua a ambiguidade verbo/interjeição na sentença *pôs, ele pôs a mão na massa*. As linhas de leitura descartadas são prefixadas com um ponto e vírgula, e os números de ID das regras são marcados em negrito.

```
"<pôs>"
        "pô" <newlex> IN
;       "pôr" V PS 3S IND VFIN REMOVE:5712
"<$,>"
"<ele>"
        "ele" PERS M 3S NOM/PIV
;       "ele" N M S REMOVE:6237
"<pôs>"
        "pôr" V PS 3S IND VFIN SELECT:5894
;       "pô" <newlex> IN SELECT:5894
"<a>"
        "o" <dem> DET F S
;       "a" PRP REMOVE:4756
;       "a" N M S REMOVE:4778
;       "ela" PERS F 3S ACC REMOVE:6729
"<mão>"
        "mão" N F S
"<em>"
        "em" <sam-> PRP
"<a>"
        "o" <-sam> <art> DET F S
"<massa>"
        "massa" N F S
"<$.>"
```

[18] E possivelmente o advérbio *pois*, que entretanto não apareceu como *pôs* no léxico de padronização.

5. A segmentação sintática

Enquanto a língua escrita oferece marcadores de parágrafo, quebras de linhas, pontos finais e outros tipos de pontuação para se deduzirem as estruturas sintática e informacional, tal segmentação é implícita ao invés de explícita na transcrição da língua falada. Em seu nível de superfície, os dados da fala não têm pontuação e, textualmente, não têm fronteiras de orações claras. Para tornar as coisas ainda piores, os dados da fala são repletos de barulho "sintático", como repetições, falsos começos e pausas ou interjeições fáticas (*ah*, *eeh*, *uh*). Entretanto, a informação marcada na transcrição nos permitiu superar a maioria desses problemas e converter os dados para um formato textual que poderia ser processado por qualquer *parser*.

A seção 3 descreveu a nossa solução para o "ruído sintático", e agora nós passaremos a enfocar a segmentação. A informação necessária para a segmentação da fala reside na prosódia (isto é, ritmo, acento e entonação) assim como nos sinais não verbais. Dependendo de se e como essa informação é codificada na transcrição, um *parser* pode simplesmente não ter a informação de segmentação necessária para funcionar adequadamente. Alguns *corpora* de fala, como a versão do *corpus* NURC descrita em Bick (1998), usa meios ortográficos para expressar a quantidade vocálica (*'u::m'*), acento (*'esnoBAR'*) e, até mesmo, pausas (*'eee'*), adicionando assim tanto mais dificuldades para o reconhecimento das palavras quanto a necessidade de inserção e desambiguação contextual de pausas, por um lado, e quebras sintáticas verdadeiras, por outro. No C-ORAL-BRASIL, ao invés de codificar informação prosódica na ortografia, a segmentação prosódica foi marcada explicitamente, no momento da transcrição, através da utilização de três padrões de segmentação distintos:

1. quebras prosódicas terminais (/ /), separando o que funcionalmente poderia ser chamado de enunciados, equivalente ao que na língua escrita seriam separações de orações;
2. quebras de descontinuação (+) entre enunciados;
3. quebras prosódicas não terminais (/), separando o que poderia ser considerado como unidades informacionais.

Ao invés de tornar essa informação invisível ao *parser* via utilização de metaetiquetas (a estratégia adotada para o ruído sintático), decidimos substituir os marcadores prosódicos por pontuação padrão, utilizando um ponto e vírgula como o equivalente mais óbvio para as quebras terminais

(//), algumas vezes o alternando com '...' para interrupções; e uma vírgula para as quebras não terminais (/). A ortografia do português não usa vírgulas obrigatórias em todos os lugares em que tínhamos a barra simples, mas a inspeção dos resultados da anotação mostrou que as vírgulas extras ajudaram ao invés de atrapalhar. Nos termos da CG, a vírgula é um membro do conjunto de BARRIER em muitas regras contextuais, separando material interno ao sintagma de itens pertencentes a outro sintagma. Sendo assim, são as regras de contexto globais (*1 e *-1) que tendem a usufruir da introdução da marcação de pontuação prosódica. Seguindo o mesmo raciocínio, a sintaxe deve ser mais afetada que a etiquetagem morfológica/PoS, uma vez que contextos mais amplos são mais importantes para a captura de relações sintáticas. O detalhamento do escopo de regras na gramática do PALAVRAS (Bick, 2000) ilustra a relativa importância do contexto global para diferentes tarefas e níveis de heuristicidade:

	Alvos morfológicos				**Alvos sintáticos**				**Todos**
	Seguros	**Níveis heurísticos**			**Seguros**	**Níveis heurísticos**			
		1.	2.	3.		1.	2.	3.	
REMOVE tag (somente contextos locais)	403	112	13	27	153	37	4	2	651
REMOVE tag (≥ 1 contextos globais)	183	44	5	5	941	219	17	1	1415
Local/global	**2.2**	2.5	2.6	5.4	**0.2**	0.2	0.2	2.0	**0.5**
SELECT tag (somente contextos locais)	271	70	8	7	60	2	1	1	420
SELECT tag (≥ 1 contextos globais)	129	23	9	2	209	57	3	-	432
Local/global	**2.1**	3.0	0.9	3.5	**0.3**	0.0	0.3	-	**1.0**

Os números mostram que a parcela de regras globais é substancial mesmo para a morfologia (cerca de 31%), e é muito alta para a sintaxe, já que nesta a maioria das regras utiliza contextos sem fronteiras. Uma tendência

geral é que regras não heurísticas tendem a ter mais contextos globais que regras em seções heurísticas. A razão pela qual até mesmo a morfologia necessita de regras globais não é apenas o tamanho e a estrutura variáveis dos sintagmas, mas também o fato de que a sintaxe – isto é, na forma como palavras funcionais e o potencial combinatório dos verbos – pode ter um papel indireto nas regras de PoS.

Pode-se concluir que para a anotação de dados de fala é imprescindível que se forneça ao *parser* algum tipo de indicação de limite relativo à oração e estrutura sintagmática, se se desejar que as regras globais funcionem otimamente. Dada a marcação de quebras prosódicas preexistente no nosso *corpus*, estas eram a escolha óbvia para candidatos à delimitação. Porém, estratégias alternativas – apesar de serem mais heurísticas – são possíveis mesmo na inexistência de tal marcação. Assim, Bick (1998) introduziu a ideia de "marcadores de *diesão*"[19] baseados em pausas, marcação de acento e interjeições de hesitação (eh, éh), que foram etiquetadas e desambiguadas como etiquetas <break> ou <pause>, em que a primeira constituiu uma quebra de frase ou oração, e a segunda é utilizada dentro de frases e até mesmo dentro de sintagmas. Os marcadores de *diesão* podem ser inseridos próximos à anotação de traços prosódicos como entonação, pausas, interjeições etc., mas podem também ser mapeados em palavras comuns, utilizando-se regras contextuais da CG para definir bordas sintáticas, tais como conjunções para orações, preposições para SPs e artigos e determinantes para SNs.

Para o *corpus* C-ORAL-BRASIL, essa técnica foi implementada explorando-se os marcadores prosódicos explícitos nele contidos. O principal marcador, correspondente à quebra terminal (/ /), foi substituído por ponto e vírgula, enquanto o marcador de quebra não terminal (/) foi reetiquetado como uma vírgula com duas leituras potenciais – <break> e <pause> –, em que somente a primeira representa uma quebra sintática, enquanto a segunda é permitida dentro de sintagmas e entre verbo e complemento. Regras da CG foram escritas para distinguir entre essas duas leituras, e a vírgula foi substituída com uma metaetiqueta para os casos de <pause>, tornando-a invisível para regras de desambiguação normais da CG. A desambiguação contextual das funções das quebras prosódicas nos permitiu achar um equilíbrio entre simplesmente ignorar a marcação, por um lado, e

[19] O neologismo diesão refere-se à tradução de *dishesion*, termo técnico utilizado principalmente na área médica, que significa separação, ruptura. (N.T.)

a supersegmentação sintática, por outro. Se fosse para as regras originais do *parser* funcionarem otimamente, então elas precisariam de uma marcação de vírgulas tão próxima da língua escrita padrão quanto possível.

Listadas a seguir estão algumas regras da CG usadas para essa tarefa de desambiguação em uma versão textualizada:

- um marcador prosódico (/) é tratado como <break> se ele ocorrer antes da primeira palavra de um SN, ou antes de um pronome na forma nominativa, seguido de um verbo finito à direita (isto é, em início de oração);
- entre um nome ou um pronome nominativo à esquerda, e um verbo finito à direita, um marcador prosódico (/) é tratado como <pause> (caso sujeito - verbo);
- marcadores prosódicos (/) entre um nome e outro SN são tratados como <break> (aposições);
- marcadores prosódicos (/) entre partes potenciais de SN são tratados como <pause> se houver concordância de gênero ou número entre os candidatos constituintes de um SN (ex.: DET-ADJ-N);
- entre um nome e um adjetivo que concordam em gênero e número, um marcador prosódico (/) é tratado como <pause> (isto é, N ADJ);
- entre um verbo transitivo e uma borda de SN à esquerda, um marcador prosódico (/) é tratado como <pause>;
- entre um auxiliar e seu verbo principal, um marcador prosódico (/) é tratado como <pause>;
- se uma única palavra é cercada por marcadores prosódicos (/), eles são tratados como <pause>;
- se um marcador prosódico (/) é precedido por uma conjunção ou relativo, ele é tratado como <pause>;
- se um marcador prosódico (/) é precedido por uma preposição, ele é tratado como <pause> (isto é, interno a um SP);
- entre um intensificador e um atributo, o marcador prosódico (/) é tratado como <pause> (isto é, interno a um SADJ);
- se um marcador prosódico (/) é seguido de certas preposições tipicamente pós-nominais (de, em, com, sem) em certos contextos, ele é tratado como <pause>.

Obviamente, uma vez que essa seção de regras teve que ser rodada *antes* que as regras do próprio *parser* o fossem (já que se destinavam a ajudá-las), as condições de contextos linguísticos tiveram que ser trabalhadas cuidadosa e não muito explicitamente, levando-se em consideração a alta ambiguidade morfológica e de PoS do insumo de texto bruto. Dentro da gramática normal do PALAVRAS, as etiquetas <break> funcionam como vírgulas, em ambas as definições estabelecidas – condições BARRIER e condições contextuais comuns.

6. A avaliação

A fim de avaliarmos o desempenho do *parser* modificado para os nossos dados, um arquivo de transcrição (*bfamdl15*) foi escolhido aleatoriamente, analisado automaticamente e corrigido manualmente. Usamos, então, a ferramenta de avaliação da CG eval_cg para comparar o arquivo da análise bruta com a versão revisada. Num uso comum da CG, a metamarcação e a pontuação estariam 100% alinhadas, mas no nosso caso, as questões foram complicadas pelo fato de que "vírgulas" haviam sido desambiguadas ou como quebra <break> ou pausa <pause>, e no último caso, foram substituídas por uma metaetiqueta. Por um lado, isso causou problemas de alinhamento para o avaliador, e por outro, as diferenças tiveram que ser identificadas e contadas como erros de cobertura. Outras incompatibilidades, causadas por divisões errôneas ou não divisões de expressões de múltiplas palavras ambíguas, também foram contadas como erros de cobertura, como por exemplo no caso de *"primeiro=que"* (conjunção *versus* adjetivo/numeral + relativo). Incluindo-se os itens de pontuação, o arquivo continha 1.895 itens.

	Cobertura	**Precisão**	**F-Score**
Função sintática	95,3	94,9	95
PoS/Classe de palavra	98,5	98,7	98,6
Morfologia	98,4	98,6	98,5
Forma base	98,6	99,4	99

Pode ser visto, através desses dados, que a tarefa mais fácil foi a lematização (formas base), enquanto a função sintática foi a mais difícil. A diferença entre cobertura e precisão para a sintaxe é uma medida de etiquetas ambíguas remanescentes. Para classes de palavra e morfologia, somente

uma leitura era permitida; dessa forma, as diferenças entre precisão-cobertura são inteiramente devidas às diferenças em alinhamento entre marcadores de quebras (vírgulas).

A fim de julgar a efetividade do uso dos marcadores de quebras prosódicas como pontuação, nós também comparamos uma rodada padrão (com desambiguação de pausa/quebra) com uma rodada sem quebras (marcadores (/) ignorados), uma rodada sem orações (tanto (/), (+) quanto (//) ignorados) e uma rodada com todas como quebras (todas as marcas (/) transformadas em vírgulas, *sem* desambiguação). Uma vez que o arquivo referência não tinha vírgulas de desambiguação, o avaliador foi rodado em modo compatibilidade somente, comparando etiquetas apenas com itens compatíveis. Dessa forma, os dados na tabela abaixo só podem ser comparados uns com os outros, e não com o teste original rodado.[20]

	Sem oração	**Sem quebra**	**Todos quebra**	**Pausa / Quebra**
Função sintática	86.2 (R: 86.5, P: 86.1)	90.7 (R: 91.0, P: 90.6)	93.7 (R: 93.3, P: 93.6)	**95.0** (R:95.3, P: 94.8)
PoS/Classe de palavra	98,3	98,8	99,3	**99,4**
Morfologia	98,1	98,6	**99**	98,7
Forma base	99	99,1	**99,4**	**99,4**

Claramente, a exploração dos marcadores prosódicos melhorou os resultados em todos os níveis. Entretanto, o efeito foi muito mais marcado para a sintaxe que para a classe de palavra, lematização e morfologia, refletindo o escopo contextual mais amplo para as etiquetas sintáticas e a resultante maior necessidade de segmentação precisa e correta.[21] Interessantemente, enquanto o desempenho sintático pode ser mais aumentado pela desambiguação pausa/quebra, isso não é óbvio para as etiquetas de

[20] Adicionalmente, esse experimento foi feito mais tardiamente quando algumas melhorias na gramática geral já haviam sido feitas, tornando uma comparação direta impossível.

[21] Conclui-se que o efeito positivo de tal segmentação em regras e não regras do princípio de unicidade (que usufruem de segmentação mais refinada) supera o bloqueamento de vírgula potencial de regras *positivas* buscando por relações sintáticas de longa distância.

categorias mais locais. Assim, para etiquetas de flexão (morfologia), o desempenho da rodada de todas as instâncias sendo consideradas como quebras foi *mais alto* que para a rodada de pausa/quebra, e somente para classe de palavra observou-se uma melhora discreta.

7. Conclusão

Enquanto o projeto de anotação do C-ORAL-BRASIL demonstrou que um *parser* criado para a língua escrita padrão (PALAVRAS) pode ser utilizado para designar etiquetas morfossintáticas para dados transcritos de fala, ele também demonstrou que para um desempenho ótimo certas adaptações devem ser feitas tanto para o sistema quanto para os dados, que incluem normalização ortográfica e extensões lexicais, assim como segmentação sintática. Esta última mostrou-se especialmente importante para a sintaxe e foi obtida através do uso dos marcadores de quebras prosódicas como "pontuação", enriquecida por distinções baseadas em regras entre funções de pausa e quebra. Sob condições ótimas, o sistema modificado do *parser* alcançou índices de acerto (F-scores) de 98,6% para classe de palavras, 95% para função sintática e 99% para lematização.

O esquema de anotação implementado preserva a informação prosódico-transcritiva original, incluindo fluxo de fala, retrações, sobreposições, tomadas de turno etc., codificadas como metaetiquetas acompanhando as etiquetas morfossintáticas, mas ficará como tarefa futura criar-se um sistema integrado de buscas (interface GUI) que permitirá ao usuário trabalhar com esses dois níveis distintos de anotação ao mesmo tempo, ao invés de separadamente. Anotação de longo termo, em altos níveis de anotação gramatical, poderia também ser adicionada e disponibilizada para buscas, tais como árvores de dependência, classes semânticas, caso e relações anafóricas; todos esses itens, em princípio, estão disponíveis para análises da língua escrita via PALAVRAS.

TRADUÇÃO DE
Heliana Mello

8. Apêndice: definições das etiquetas

8.1 Verbos

V verbo

pessoa/número

1S	primeira pessoa do singular
2S	segunda pessoa do singular
3S	terceira pessoa do singular
1P	primeira pessoa do plural
2P	segunda pessoa do plural
3P	terceira pessoa do plural

0/1/2/3S infinitivo morfema zero

modo

IND	indicativo
SUBJ	subjuntivo
COND	condicional
IMP	imperativo

tempo

PR	presente
IMPF	passado (imperfeito)
FUT	futuro (simples)
PS	passado (perfeito simples)

Formas não finitas

INF	infinitivo (+ número/gênero)
GER	gerúndio
PCP	particípio (+ número/gênero)

8.2 Nominais e pré-nominais

N	nome
PROP	nome próprio
ADJ	adjectivo
DET	determinante
NUM	numeral

gênero

M	masculino
F	feminino
M/F	invariável/ subespecificado

número

S	singular
P	plural
S/P	invariável/ subespecificado

Classes nominais secundárias

<KOMP>	comparativo
<SUP>	superlativo
<NUM-ord>	ordinal
<card>	cardinal
PERS	Pronome pessoal

pessoa/número

1S	primeira pessoa do singular
2S	segunda pessoa do singular
3S	terceira pessoa do singular
1P	primeira pessoa do plural
2P	segunda pessoa do plural
3P	terceira pessoa do plural

gênero

M	masculino
F	feminino
M/F	invariável/subespecificado

caso

NOM	nominativo
ACC	acusativo
DAT	dativo
PIV	prepositivo
NOM/PIV	subespecificado
ACC/DAT	subespecificado
INDP	Independente (não flexional)

Classes secundárias de pronomes

<arti>	DET	artigo indefinido
<artd>	DET	artigo definido
<dem>	DET, INDP	demonstrativo
<poss...>	DET	possessivo
<quant>	DET	quantificador
<rel>	DET, INDP	relativo
<interr>	DET, INDP	interrogativo
<refl>	PERS	pronome pessoal reflexivo
<si>	DET	possessivo reflexivo

8.3 Classes de palavras não flexionais

ADV advérbio

<rel> relativo
<interr> interrogativo
<ks> semelhante a conjunção subordinativa
<kc> semelhante a conjunção coordenativa
<foc> marcador de foco

PRP proposição
KS conjunção subordinativa
KC conjunção coordenativa
IN interjeição

8.4 Etiquetas sintáticas

@SUBJ> @<SUBJ sujeito

@ACC> @<ACC objeto (direto) acusativo

@DAT> @<DAT objeto dativo (somente pronominal)

@PIV> @<PIV objeto (indireto) preposicionadao

@ADVS> / @SA> @<ADVS / @<SA predicativo adverbial (lugar, tempo, duração, quantidade) equivalente a nome @SC

@ADVO> / @OA> @<ADVO / @<OA objeto predicativo adverbial, equivalente a nome @OC

@SC> @<SC predicativo do sujeito

@OC> @<OC predicativo do objeto

@ADVL> @<ADVL expressão adverbial

@PASS> @<PASS agente da passiva

@ADVL sintagma adverbial "livre" (em expressão não sentencial)

@NPHR sintagma nominal "livre" (em expressões não sentenciais, sem verbos)

@VOK "vocativo" (ex. "livre" referindo-se a nome próprio em discurso direto)

@>N adjeção pré-nominal

@N< adjeção pós-nominal

@N<PRED pós-nominal (dentro de um predicativo)

@APP aposição de identificação

@>A adjeção adverbial preposta

@A< adjeção adverbial posposta

@PRED> predicativo livre "avançado"

@<PRED predicativo livre "atrasado"

@P< argumento de preposição

@S< aposição relacionada à sentença

@FAUX auxiliar finito (cp. @#ICL-AUX<)

@FMV verbo principal finito

@IAUX auxiliar não finito (cp. @#ICL-AUX<)

@IMV verbo principal não finito

@PRT-AUX< particular da cadeia verbal (preposição ou "que" após auxiliar)

@CO conjunção coordenativa

@SUB conjunção subordinativa

@KOMP< argumento de comparativo (ex. "do que" referindo-se a *melhor*)

@COM comparador direto sem comparativo

@PRD predicador de papel/função (ex. "trabalha *como*")

@FOC> @<FOC marcador de foco ("gosta *é* de peixe.")

@TOP constituinte de tópico ("*Esse negócio*, não gosto dele.")

@#FS- suboração finita (combina-se com papel da oração e etiqueta para palavra intrassentencial, ex. @#FS-<ACC @SUB para "não acredito *que* seja verdade")

@#ICL- suboração não finita (combina-se com papel sentencial e etiqueta de palavra intrassentencial, ex. @#ICL-SUBJ> @IMV em "*consertar* um relógio não é fácil")

@#ICL-AUX< verbo argumental numa cadeia verbal, refere-se ao auxiliar precedente

@#AS- suboração averbal (isto é, sem verbo) (combina-se com papel sentencial e etiqueta de palavra intrassentencial, ex.@#AS-<ADVL @ADVL> em "ajudou *onde* possível")

@AS< argumento do complementizador em suboração averbal

8.5 Algumas etiquetas secundárias

<*>	maiúscula
<*1>	aspas à esquerda
<*2>	aspas à direita
<sam->	primeira parte de contração
<-sam>	segunda parte de contração
<parkc-1>	primeira parte em *ou...ou*
<parkc-2>	segunda parte em *ou...ou*
<pp>	advérbio complexo (sintagma preposicionado)
<hyfen>	palavra hifenizada
<newlex>	originário do léxico adicional

O PALAVRAS também utiliza cerca de 200 etiquetas prototípicas para nomes, não listadas aqui, assim como etiquetas de valências para verbos, nomes e adjetivos.