Tratamento dos sufixos modo-temporais na depreensão automática da morfologia dos verbos do português

Vera Vasilévski
Universidade Federal de Santa Catarina
sereiad@hotmail.com

Márcio José Araújo
Universidade Tecnológica Federal do Paraná
marciomjapr@gmail.com

Resumo

Este artigo apresenta um analisador morfológico automático de verbos do português, com destaque para seu desempenho no processamento das regras que regem esse sistema verbal e no tratamento das ambigüidades geradas. Nesta etapa, trabalha-se com as ambigüidades decorrentes da alomorfia dos sufixos modo-temporais e da possibilidade de esses morfemas serem zero (Ø) em alguns modos e tempos, nas três conjugações do português. Para esclarecer o trabalho feito com o analisador, traz um resumo das regras morfológicas do sistema de verbos do português. Obteve-se êxito no tratamento de muitas das ambigüidades que o programa registrou, as quais eram esperadas, uma vez que coincidem com as ambigüidades do sistema de verbos da língua portuguesa. A resolução da maioria delas fez-se com base em regras computacionais (estruturas de seleção) que consideram o contexto do enunciado. Conclui que a resolução de outras ambiguidades relacionadas a modo e tempo verbal somente será possível ao se levar em conta também os morfemas número-pessoais, que são objeto de outro trabalho.

1. Introdução

Os computadores e, a partir deles, a Lingüística Computacional tornaram possíveis o armazenamento e a análise de quantidade nunca antes conhecida de dados da comunicação verbal, em tempo realmente curto. Isso possibilita descrições, comparações e generalizações com base em uma massa de dados bastante densa. No entanto, apesar do desenvolvimento computacional voltado para trabalho com língua escrita e falada que se constata atualmente, ainda é reduzido o número de aplicativos dessa natureza disponíveis e efetivamente utilizados por usuários, por diversos motivos. Em especial, a metalinguagem vê-se carente de recursos eletrônicos específicos para auxiliar pesquisas em várias áreas da Lingüística.

Dentre os motivos para o pouco uso desses aplicativos, além da falta de divulgação, está a dificuldade do usuário de utilizar esses programas, às vezes, porque não consegue compreender como eles funcionam. Agrava essa situação o fato de muitos desses aplicativos serem em inglês ou feitos para essa língua, o que reduz sua eficiência para o português (Vasilévski, 2010). Ainda, não raro, seu nível de interatividade é precário (Vasilévski, 2010), o que desestimula seus potenciais usuários. Faz-se necessário remediar essa situação.

Este estudo apresenta um recurso computacional especialmente desenvolvido para automatizar o sistema de verbos do português, a partir de suas regras morfológicas, e debate algumas implicações advindas dessa tarefa. Ressalta a automatização dos morfemas modo-temporais, os casos ambíguos dela decorrentes e discute sua desambiguação. Essa ferramenta foi desenvolvida como parte do projeto *Análise morfológica Automática do Português* (Scliar-Cabral, 2009), cujo objetivo maior é depreender uma gramática automática do português brasileiro, mediante análise do córpus pau003.cha – que é constituído por 10688 enunciados e está disponível para ser baixado (Childes, 2011) –, levando-se em conta a fala dos adultos quando conversam entre si, a fala dirigida à criança e a fala da criança ao se comunicar com os adultos. Desenvolve-se esse projeto em parceria com o projeto Childes (MacWhinney, 2011). Etapas desse projeto têm sido apresentadas (Vasilévski 2010, 2011a, 2011b, 2011c; Scliar-Cabral e Vasilévski, 2011; Scliar-Cabral, 2011), e a que se expõe aqui se refere à análise morfológica automática de verbos conjugados,

portanto, em situações de uso, encontradas em um enunciado.

Buscas foram feitas em bases de artigos científicos e diretamente na Rede Mundial de Computadores, na tentativa de encontrar documentos sobre outros analisadores morfológicos de verbos do português similares a este que ora se apresenta, para aprimorá-lo, bem como para compará-los, mas não se obteve sucesso. Isso não significa que não existam trabalhos dessa natureza, no entanto, eles não estão disponíveis ou suficientemente divulgados.

A automatização que se demonstra nesta ocasião restringe-se aos verbos regulares, mas a finalidade da ferramenta computacional é analisar morfologicamente também os verbos irregulares do português. É possível testar-se qualquer forma verbal no programa, pois todos os tempos verbais e pessoas gramaticais da língua portuguesa foram inseridos em seu algoritmo. Para dar suporte às etapas do projeto, criou-se um programa que abriga várias ferramentas, as quais funcionam em conjunto e em interface com outros aplicativos. Esse programa chama-se Laça-palavras (Vasilévski e Araújo, 2010), e o analisador morfológico enfocado aqui é uma de suas ferramentas.

A partir disso, este artigo aborda o referencial teórico básico utilizado para se compreender o sistema de verbos do português e suas regras, o programa Laça-palavras, os princípios de funcionamento do analisador morfológico, suas principais convenções e seu desempenho.

2. O sistema de verbos do português

Para se desenvolver o analisador morfológico automático para verbos do português, foi necessário conhecer a fundo as regras gramaticais que regem o sistema de verbos dessa língua, o que se obteve na literatura pertinente ao tema (Câmara Jr., 1986, 1976; Scliar-Cabral, 2003, 2007, 2008). Parte dessa literatura já foi discutida (Vasilévski, Scliar--Cabral e Araújo, 2012), quando se tratou especificamente do comportamento da vogal temática, mas cabe revisitá-la e complementá--la com teoria específica que fundamente o assunto deste artigo.

O analisador morfológico em foco baseia-se em regras gramaticais, e não em aprendizado de máquina, ou seja, não gera regras automaticamente, a partir de um dicionário de treino. Ocorre sim que regras gramaticais foram convertidas em algoritmos e testadas no córpus. O projeto é criado e coordenado por cientistas da língua, e conta com o suporte indispensável da computação.

O sistema de conjugação de verbos do português é considerado, de certa forma, simples e previsível (Câmara Jr., 1986), o que respalda a criação de uma ferramenta computacional baseada em suas regras. O sistema de verbos do português compreende três conjugações, assinaladas pela vogal temática. Há três vogais temáticas, conforme a notação escrita usual: "a" para a primeira conjugação (1.ªC), "e" para a segunda conjugação (2.ªC) e "i" para a terceira conjugação (3.ªC). Compõem esse sistema verbos ditos regulares – que seguem o paradigma fixo da conjugação a que pertencem e são maioria em português – e os irregulares – que se desviam do paradigma regular.

O tema do infinitivo é a forma básica do verbo regular. Assim, dado um verbo regular em sua forma infinitiva, é possível conjugá-lo com facilidade, nas seis pessoas gramaticais, sobretudo nos tempos do modo indicativo. Em contrapartida, tomar uma forma verbal conjugada e dela extrair os morfemas que a compõem, a fim de desvendar tempo, modo, pessoa e número em que está flexionada, não é tão fácil.

Em português, há três modos verbais finitos, com seus tempos simples (indicativo (seis tempos), subjuntivo ou conjuntivo (três tempos) e imperativo (afirmativo e negativo), além do infinitivo pessoal e das formas nominais infinitivo, gerúndio e particípio. Na seção 4, expõem-se tais tempos, da maneira como foram inseridos no algoritmo morfológico. Cabe destacar que o pretérito mais-que-perfeito é pouco usado no Brasil. Na fala coloquial, ele restringe-se a frases feitas, e é raramente usado na língua escrita. Também, é preservado, sobretudo, na literatura e em músicas, e aparece esporadicamente no falar jornalístico, por exemplo, em editoriais. Quanto às formas nominais, o infinitivo é a forma mais genérica do verbo, que de maneira

mais ampla e vaga resume sua significação, sem noções de tempo, modo e aspecto. Por sso, ele é usado para designar o nome do verbo, e pode funcionar como substantivo.

Entre o gerúndio e o particípio há oposição de aspecto, pois o primeiro é imperfeito (processo inconcluso) e o segundo é perfeito (processo concluso). Morficamente, o particípio desvia-se da natureza do verbo, pois pode tornar-se um adjetivo verbal. É a única forma verbal que assume gênero e número, além da categoria de voz passiva. Assim, morfologicamente, ele pertence aos adjetivos, embora tenha valor verbal no âmbito sintático e semântico (Câmara Jr., 1986). Como verbo, o particípio entra na formação dos tempos compostos com o auxiliar "ter" e "haver" (quando permanece invariável em gênero e número) e com o auxiliar "ser", na construção da voz passiva analítica, além de núcleo do predicado de orações reduzidas. Já o gerúndio é morfologicamente verbal, assim, não admite flexão de gênero e número (Câmara Jr., 1986), e entra na formação das formas progressivas e também como núcleo do predicado de orações reduzidas.

É válido mencionar que, por irregulares, entendem-se os verbos cujos temas das formas primitivas são distintos entre si – essa é a principal característica de sua irregularidade. São formas primitivas os temas da primeira pessoa do singular e as segundas pessoas do presente do indicativo; o tema da segunda pessoa do singular do pretérito perfeito do indicativo; e o tema do infinitivo impessoal, os quais dão origem aos outros tempos verbais. Por exemplo, o tema da primeira pessoa do singular do presente do indicativo dá origem ao presente do subjuntivo. O verbo "ser", na primeira pessoa do singular do presente do indicativo, tem o tema *so-*, enquanto, na segunda pessoa do singular do pretérito perfeito do indicativo, seu tema é *fo-*, portanto, ele é irregular. Ainda, além das formas primitivas distintas, o verbo "ser" apresenta irregularidades nas derivações, como no presente do subjuntivo, cujo tema é *sej-*, em todas as pessoas. No entanto, os verbos irregulares não o são em todos os tempos e pessoas gramaticais. Nos tempos futuros (do presente e do pretérito) do modo indicativo, há pouquíssima irregularidade. Nesses tempos, verbos irregulares como "ser", "estar", "vir" são perfeitamente analisados pelo programa, como formas regulares. A exceção fica por conta dos verbos "fazer" e "trazer", cujas formas nesses tempos, para serem regulares, deveriam ser *"fazerá", *"trazerá" e *"fazeria", *"trazeria", as quais são incorretas.

Do grupo dos verbos irregulares fazem parte os verbos auxiliares, que figuram nas conjugações compostas. Em uma cadeia podem entrar vários verbos auxiliares, sendo o último verbo o verbo principal, aquele que carrega a significação externa, sempre em uma forma nominal (infinitivo, gerúndio ou particípio). É ele quem dita a regência, por isso, o primeiro auxiliar da cadeia se flexionará em pessoa, número, tempo e modo, conforme tal verbo principal determinar. Por exemplo: "**ia** entrar", "**deve estar** havendo muitas suspeitas", "**podem**-se esperar vitórias" e "**tinha** feito".

O verbo é, em português, o vocábulo flexional por excelência, dada a complexidade e a multiplicidade de suas flexões. As noções gramaticais de tempo e modo e de pessoa e número que a forma verbal indica correspondem a duas desinências (sufixos flexionais) chamadas de sufixo modo-temporal (SMT) e sufixo número-pessoal (SNP), que se aglutinam e se ligam ao tema. O tema constitui-se do radical seguido da vogal temática da conjugação correspondente. No padrão geral, o radical é invariável e dá a significação lexical do verbo. Assim, a fórmula geral da estrutura do vocábulo verbal português – na qual RAD indica radical do verbo; VT, vogal temática; e SF, sufixos flexionais – é (Câmara Jr., 1986):

TEMA (RAD+VT) + SF (SMT + SNP)

Levando-se em conta a alomorfia de cada um dos sufixos flexionais e a possibilidade de ser zero (Ø) para um deles ou ambos, essa fórmula dá a regra geral da constituição morfológica do verbo em português, além de indicar a ordem obrigatória dos morfemas. A aglutinação em um único morfema das noções de tempo e modo determina, evidentemente, 13 morfemas modo-temporais, nos quais só esporadicamente ocorre alomorfia, isto é, a variação de um morfema condicionada pelo contexto onde ele

ocorre. No analisador, são levados em conta apenas os alomorfes do sistema escrito.

A complexidade para a interpretação do morfema flexional propriamente verbal, em português, ou seja, o modo-temporal, decorre, em primeiro lugar, justamente da cumulação dessas duas noções, além da noção suplementar de aspecto, que às vezes se inclui na noção de tempo (Câmara Jr., 1986). De maneira muito resumida, pois o assunto é complexo, o tempo verbal refere-se ao momento de ocorrência do processo, visto do momento da comunicação; já o modo refere-se a um julgamento implícito do falante a respeito da natureza, subjetiva ou não, da comunicação que faz. No entanto, é comum em português, assim como em outras línguas, o emprego modal dos tempos verbais, que já foi chamado de metafórico. Não obstante, a apreciação do modo em português tem de se firmar, inicialmente, nas formas modais propriamente ditas, deixando à margem o emprego metafórico dos tempos (Câmara Jr., 1986).

Outrossim, há 06 sufixos número-pessoais, para indicar os falantes (1.ª pessoa do discurso), os ouvintes (2.ª pessoa do discurso) e as entidades sobre quem se fala (3.ª pessoa do discurso) (Câmara Jr., 1986). No português do Brasil (PB), a segunda pessoa do discurso pode se valer da terceira pessoa gramatical. As pessoas gramaticais são designadas por 1, 2 e 3 do singular (S) e do plural (P), assim, tem-se: 1S (eu), 2S (tu), 3S (ele, ela, você), 1P (nós), 2P (vós) e 3P (eles, elas, vocês). Como visto, no PB, usam-se conjugadas como 3S e 3P "você" e "vocês", respectivamente, o que aumenta o nível de ambigüidade do sistema de verbos, pois apesar das flexões da terceira pessoa, essas formas referem-se à segunda pessoa do discurso. Ainda, 1P (nós) pode ser substituída por "a gente", quando então assume as flexões de 3S ou, muito mais raramente, 1P. Os últimos casos são tão potencialmente ambíguos, que as pessoas gramaticais sempre estão expressas, o que facilita a desambiguação pelo contexto escrito.

Estudo recente (Scliar-Cabral, 2008) propõe o refinamento da fórmula anterior para:

TEMA (RAD+VT) + SF (SMT**A** + SNP + **SPF**)

com a inclusão do acento ou suprafixo (SPF) e da categoria de aspecto (A). Essa inclusão do acento de intensidade com a função de assinalar diferenças aspectuais tem sido negligenciada na literatura sobre aquisição da linguagem, o que causa problemas para a codificação automatizada ainda não tratados, como a queda do morfema *-r* do infinitivo na pronúncia. Contudo, a ferramenta que aqui se expõe lida com a língua escrita, não foca, por enquanto, esse ponto. Estima-se abordar essa questão com apoio da fonologia, em outra fase do projeto, tarefa para a qual o programa Laça-palavras já está preparado.

De posse desse conhecimento, levaram-se em conta todas essas considerações e transformaram-se essas regras – e outras não detalhadas aqui – em algoritmos. Para tanto, fizeram-se ajustes e complementações que o ambiente computacional exige, obviamente, e isso implicou a criação de novas regras. Depois, estudou-se o comportamento do aplicativo, a fim de se observarem ambigüidades geradas e resolvê-las, bem como para criar um léxico verbal automático para o córpus de trabalho. A criação do léxico – que já foi demonstrada (Vasilévski, Scliar-Cabral e Araújo, 2012) – resolveu as ambigüidades geradas pela alomorfia da vogal temática, nas três conjugações, e pela harmonia vocálica que ocorre no radical de verbos da 3.ªC, uma vez que a harmonia vocálica que ocorre no radical de verbos da 1.ªC e 2.ªC conjugações não é registrada no sistema escrito. Ainda, esse léxico, associado a instruções computacionais, resgata radicais regulares que sofrem transformações ditadas pelos valores grafêmicos, pelos quais: "g", quando vem antes de "e" e "i", é escrito "gu", para preservar o valor de /g/, como "**lig**ar" → "**ligu**ei"; "c", antes de "e" e "i", é escrito "qu", para preservar o valor de /k/, como "**fic**ar" → "**fiqu**ei"; e "c", antes de "o", "a" e "u", é escrito "ç", para preservar o valor de /s/, como "**esquec**er" → "**esqueç**o", "**esqueç**a". Trabalho a ser publicado detalha esse processo e outros semelhantes.

Antes de passar-se à ferramenta para análise morfológica dos verbos do português, cabe resumir o funcionamento e os recursos do programa Laça-palavras, que é o ambiente no qual está inserida essa ferramenta.

3. O programa Laça-palavras

O funcionamento geral do programa Laça--palavras foi relatado anteriormente (Scliar--Cabral e Vasilévski, 2011), bem como algumas de suas ferramentas (Vasilévski, 2010, 2011a, 2011b, 2011c) e resultados oriundos de sua implementação parcial (Vasilévski, 2011d; Costa e Scliar-Cabral, 2011).

O Laça-palavras (LP) surgiu da necessidade de haver flexibilidade dos dados de trabalho maior do que a oferecida pelo programa Clan, disponibilizado pela Plataforma Childes e usado para se ler o córpus. Foi preciso se disporem os dados de diferentes formas e se extraírem deles informações que não eram possibilitadas pelo Clan. O Laça-palavras volta-se para arquivos em português, trabalha em conjunto com o Clan e também disponibiliza recursos próprios.

As interfaces do Laça-palavras com o programa Clan e as diretrizes dessa interação já foram descritas (Scliar-Cabral e Vasilévski, 2011). Então, cabe apenas lembrar suas principais ferramentas: 1) pesquisa no córpus, com marcação das linhas de seus enunciados com o tipo de discurso – de adulto para criança (ad-chi), de criança para adulto (chi-ad) e de adulto para adulto (ad-ad) –, resgate de palavras específicas – ou grupos de palavras – para trabalho com classes gramaticais, geração de relatório estatístico; 2) criação no córpus de uma linha denominada %pho, mediante interface com o programa Nhenhém (Vasilévski, 2008), para fazer a transcrição fonológica automática, com marcação das sílabas tônicas de determinado enunciado do arquivo, com ajuste da transcrição fonológica para fonética; 3) criação de uma linha para tradução morfológica automática dos verbos, chamada %mor, cuja ferramenta que a controla é foco deste estudo. Apesar de estar dentro do Laça-palavras e de isso facilitar sobremaneira seu uso, o analisador morfológico, quando estiver concluído em todas suas etapas, poderá ser instalado diretamente no computador, sem obrigatoriedade de haver também instalado o Laça-palavras. Não obstante, a integração entre ferramentas traz vantagens ao usuário do analisador, já que elas se comunicam entre si e compartilham resultados.

O processamento automático das unidades morfológicas dos enunciados do córpus coloca à disposição dos pesquisadores que trabalham com a morfologia do português uma eficiente ferramenta para análises quantitativas e qualitativas. No plano teórico, contribui em nível explicativo para melhor compreensão da construção das gramáticas do PB, particularmente, do sistema de verbos, e amplia o entendimento sobre o papel do *input* na construção de tais gramáticas (Scliar--Cabral, 2008), além de demonstrar a intuição do adulto, ao utilizar um registro adequado ao nível da criança.

4 Padrões e convenções do analisador

O correto funcionamento do programa depende da metodologia empregada, sobretudo, na delimitação das tarefas que ele deve executar e na criação de códigos para as categorias que ele deve controlar. O analisador está preparado para carregar e ler arquivos criados no programa Clan, então, adotaram-se convenções estipuladas por esse programa para as classes gramaticais e para anotar córpus de língua oral (MacWhinney, 2000), bem como se criaram outras convenções específicas para o analisador morfológico.

4.1 Preparação do córpus

Para a pesquisa, mostrou-se relevante anotar os verbos diretamente no córpus, na linha do enunciado, no sistema Clan, com @v (verbos regulares – *default*), @vi (verbos irregulares) e @va (verbos auxiliares), para possibilitar, no Laça-palavras, a pesquisa (resgate e filtragem de dados), a análise morfológica automática e, conseqüentemente, a criação da linha %mor no arquivo original a ser lido pelo Clan. No entanto, para fins de clareza e limpeza do texto, esses símbolos, bem como outros símbolos do Clan, podem ser omitidos na pesquisa feita pelo LP, a critério do usuário. Todos os verbos auxiliares são irregulares, mas a decisão de assinalá-los separadamente se deve ao fato de preparar a computação, posteriormente, das locuções verbais e dos tempos compostos (Scliar-Cabral e Vasilévski, 2011).

4.2 Nomenclatura

Além da notação no córpus com @v, @vi e @va, usou-se um código para cada um dos tempos verbais do português, em seus respectivos modos. Assim, inseriram-se no programa os seguintes códigos: PI – Presente do Indicativo, PII – Pretérito Imperfeito do Indicativo, PPI – Pretérito Perfeito do Indicativo, PMI – Pretérito Mais-que-perfeito do Indicativo, FPI – Futuro do Presente do Indicativo, FPPI – Futuro do Pretérito do Indicativo, PS – Presente do Subjuntivo, PIS – Pretérito Imperfeito do Subjuntivo, FS – Futuro do Subjuntivo, IMA – Imperativo Afirmativo, IMN – Imperativo Negativo, INF – Infinitivo, GER – Gerúndio, PAR – Particípio.
Da mesma forma, as pessoas gramaticais, como visto, são assim designadas: 1S, 2S, 3S, 1P, 2P e 3P.

5 Análise morfológica automática

Como mencionado, para a criação da linha %mor, foi desenvolvida uma ferramenta específica, o analisador, cujo algoritmo contém as regras das três conjugações verbais, em seus respectivos modos e tempos (Vasilévski, 2011b). Cabe esclarecer que o sistema não conjuga verbos, mas sim analisa entradas, que devem ser formas verbais escritas corretamente flexionadas.

5.1 Regras gerais

O primeiro conjunto de regras gramaticais desenvolvido foi relativo às vogais temáticas, seguido das regras dos morfemas modo-temporais e então das regras dos morfemas número-pessoais, para os verbos regulares. Tais regras foram formalizadas, para posterior inserção no programa. As regras da vogal temática foram objeto de trabalho anterior (Vasilévski, Scliar-Cabral e Araújo, 2012) e as regras específicas das pessoas gramaticais serão objeto de trabalho futuro. Aqui, cabe detalhar o segundo conjunto, ou seja, os sufixos ou desinências de modo e tempo, que se aglutinam em português. Exemplificam-se algumas dessas regras.

SMT	se realiza	como	em contexto	Exemplos
|-va-|	→	-ve- / -va-	a__i / ...	cantá**ve**is louva**va**, ligá**va**mos

Figura 1: Esquema de regras alomórficas do sufixo flexional modo-temporal da 1.ª C para o pretérito imperfeito do modo indicativo.

O esquema da Figura 1 mostra as regras alomórficas da desinência modo-temporal do pretérito imperfeito do modo indicativo, para a 1.ªC, no qual se nota que somente há alomorfia (de *-va-* para *-ve-*) na segunda pessoa do plural (vós), na qual o SMT está entre as vogais "a" e "i", contexto que condiciona a alomorfia. Nas demais pessoas (...), permanece o SMT inicial. Aliás, a forma pessoal "vós" tem uso restrito no Brasil, mas é usada em outros países em que se fala português. Preservam-se no programa formas pouco usadas no Brasil – por estarem consagradas na literatura e ainda em uso no discurso atual religioso, bem como nas músicas desse teor, por exemplo –, pois um sistema automático deve abranger todas as possibilidades oferecidas pela língua, sejam elas pouco ou muito usadas, e isso vale para os tempos verbais.

SMT	se realiza	como	em contexto	Exemplos
|-ia-|	→	-ie- / -ia-	__i / ...	venc**ie**is aplaud**ia**m

Figura 2: Esquema de regras alomórficas do sufixo flexional modo-temporal da 2.ªC e 3.ªC para o pretérito imperfeito do modo indicativo.

O esquema da Figura 2 mostra que, no pretérito imperfeito do modo indicativo, na 2.ªC e 3.ªC, somente há alomorfia em 2P.
O esquema da Figura 3, a seguir, mostra que, na 1.ªC, 2.ªC e 3.ªC, no futuro do presente do modo indicativo, o morfema respectivo *-re-* sofre alomorfia para *-rá-*, em fim de vocábulo (#) e antes de "s" em fim de vocábulo, ou seja, em 2S e 3S, e sua vogal aberta "a" recebe til diante da vogal "o" em fim de vocábulo, ou seja, em 3P. Da mesma forma, a partir do esquema da Figura 4, observa-se que, no futuro do pretérito do modo indicativo, para as três

conjugações do português, somente há alomorfia da desinência *-ria-* em 2P, ou seja, diante da vogal "i".

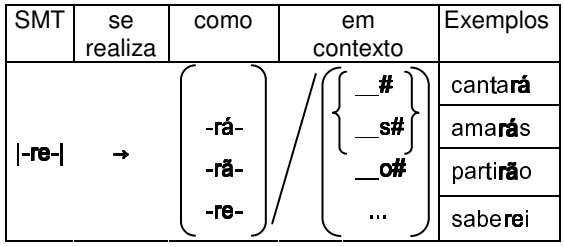

Figura 3: Esquema de regras alomórficas do sufixo flexional modo-temporal da 1.ªC, 2.ª C e 3.ªC para o futuro do presente do modo indicativo

O analisador contém as regras dos casos em que acentos gráficos ocorrem na vogal temática ("lig**á**vamos") e nas desinências, como mostram a figuras anteriores. Então, se o usuário omitir o assento gráfico do vocábulo verbal a ser analisado, o sistema poderá acusar erro, por não encontrar uma regra em que tal vocábulo se encaixe, ou encaixá-lo em uma regra incorreta.

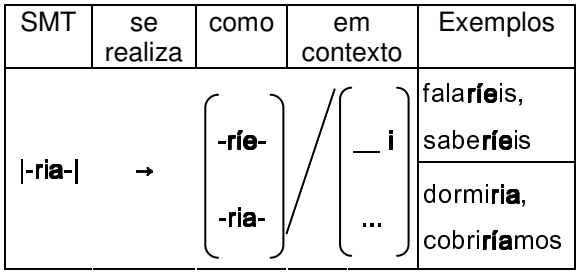

Figura 4: Esquema de regras alomórficas do sufixo flexional modo-temporal da 1.ªC, 2.ª C e 3.ªC para o futuro do pretérito do modo indicativo.

A partir das regras formalizadas e com apoio da literatura, fez-se um quadro geral do comportamento dos morfemas modo-temporais, com suas respectivas alomorfias, em parênteses, para os tempos verbais do português:

Quadro 1: Regras alomórficas dos sufixos modo-temporais do português.

MT	**SMT**			Onde há alomorfia
	1.ª C	**2.ª C**	**3.ª C**	
PI	Ø	Ø	Ø	-
PII	va (ve)	ia (ie)	ia (ie)	2P
PPI	Ø (ra)	Ø (ra)	Ø (ra)	3P
PMI	ra (re)	ra (re)	ra (re)	2P
FPI	re (rá, rã)	re (rá, rã)	re (rá, rã)	2S,3S,3P
FPPI	ria (ría,ríe)	ria (ría, ríe)	ria (ría, ríe)	1P e 2P
PS	e	a	a	-
PIS	sse	sse	sse	-
FS	r (re)	r (re)	r (re)	2S
IMA	e (Ø)	a (Ø)	a (Ø)	2S e 2P
IMN	e	a	a	-
INF	r (re)	r (re)	r (re)	2S
GER	ndo	ndo	ndo	-
PAR	do	do (to)	do	-

Cabe destacar que o pretérito imperfeito do subjuntivo e o gerúndio são tempos não ambíguos, pois seus morfemas são exclusivos e não há alomorfes. Os futuros do presente e do pretérito do indicativo têm alomorfes, os quais são exclusivos desses tempos, o que também torna esses tempos não ambíguos. Os demais tempos estão sujeitos a ambigüidades em alguma conjugação. Esse assunto voltará a foco.

5.2 Regras dos verbos irregulares

O algoritmo que contém as regras verbais está em fase de aprimoramento, para que dê conta dos verbos irregulares do PB. Assim, é válido esboçar algumas diretrizes que guiarão tal trabalho.

Verbos irregulares, na verdade, são formas irregulares, que devem ser entendidas como desvios do padrão geral morfológico, que não deixam de ser regulares, no sentido de que são suscetíveis a uma padronização. Trata-se de pequenos grupos de verbos, com certos padrões comuns, que podem ser explicitados (Câmara Jr., 1986). Tais irregularidades podem ser referir aos sufixos, mas, quando ocorrem no radical, são muito mais relevantes para a análise morfológica automática, pois se cria uma série de padrões morfológicos. Ainda, nesses verbos ocorre constantemente a supressão da vogal temática, o que acontece também na segunda e terceira conjugações com os verbos regulares, e provoca entrave na análise morfológica automática, pois se perde a conjugação do verbo, o que conseqüentemente dificulta o resgate da forma infinitiva desse verbo. A partir disso, entende-se que poderá ser bem-sucedida a decomposição morfológica automática desses verbos.

6. Desempenho

O analisador morfológico verifica os verbos contidos em um córpus previamente preparado, carregado no sistema Laça-palavras, que o abriga. As formas verbais anotadas são automaticamente lidas e analisadas, e o resultado é mostrado. Internamente, ocorre que, ao identificar uma forma verbal, o analisador morfológico a compara com suas regras internas, para decompô-la em morfemas.

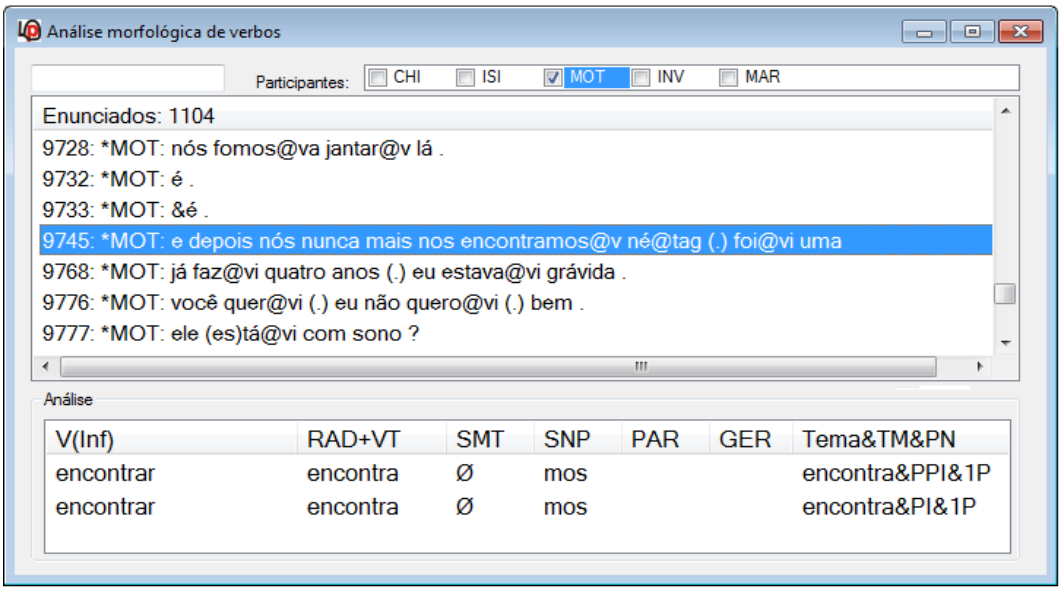

Figura 5: Tela principal do analisador morfológico automático.

Assim, a saída do programa é a realização da fórmula geral da estrutura do vocábulo verbal português e a depreensão de sua forma infinitiva. Cabe esclarecer que a criação do léxico dos verbos do córpus resolveu problemas de sobregeração de formas infinitivas, causada por conflito entre as regras do sistema de verbos do português. Por exemplo, a entrada "cantava" gerava as opções de formas infinitivas "cantar", *"cantavar", *"cantaver" e *"cantavir", das quais somente a primeira é correta e existe em português. Com a criação do léxico, o programa faz a análise e obtém todas as formas possíveis, mas, antes de mostrá-las, compara-as com o conteúdo do léxico verbal. Então, escolhe a forma que está contida no léxico e mostra somente ela (Vasilévski, Scliar-Cabral e Araújo, 2012).

A Figura 5 traz o resultado da análise morfológica da forma verbal "encontramos", que ocorre no enunciado da linha 9745 do córpus – o arquivo pau003.cha – e é proferido pelo participante MOT (a mãe da criança), quando se dirige a outro adulto. À medida que o cursor desce pelos enunciados, cada forma verbal identificada é analisada automaticamente pelo programa. Ao encontrar mais de um verbo na mesma linha do enunciado – por exemplo, na linha 9728, em que há "fomos" e "jantar" –, o programa analisa-o abaixo do verbo anterior. O campo Participantes permite ao usuário escolher os participantes cujos enunciados ele quer verificar. No córpus, há cinco participantes, sendo CHI a criança, os demais são adultos. Ao carregar o arquivo, o programa preenche esse campo com os participantes automaticamente, para o usuário selecionar os que deseja checar. Após essa seleção, aparecerá, no campo Enunciados, a quantidade de enunciados encontrados referentes ao(s) participante(s) selecionado(s), e abaixo, os enunciados propriamente ditos, no formato do Clan.

Para verificar os verbos de um enunciado, o usuário clica sobre ele, e o campo Análise fornecerá a análise morfológica do verbo que aparece no enunciado selecionado. A primeira informação morfológica fornecida é a forma infinitiva do verbo em questão (V(Inf)), seguida pela depreensão do tema (RAD+VT), desinências modo-temporal (SMT) e número-pessoal (SNP), formas nominais particípio (PAR) e gerúndio (GER). Finalmente, essa análise morfológica é traduzida em uma

seqüência de morfemas separados pelo caractere & – que nesse caso indica inserção de sufixo –, a qual repete o tema e fornece tempo/modo e pessoa/número verbais: Tema&TM&PN.

Assim, a Figura 5 mostra que a forma verbal "encontramos" é da 1.ªC, pois sua forma infinitiva (Inf) é "encontrar" e sua vogal temática é "a", que, portanto, não sofreu alomorfia nem desapareceu; não há morfema modo-temporal, o que é assinalado por Ø; o sufixo número-pessoal embutido nela é "mos"; e não se trata de particípio nem de gerúndio. Tudo isso junto diz que essa forma verbal está conjugada em 1P do presente ou do pretérito perfeito, ambos do modo indicativo.

Observam-se, nos enunciados que aparecem na Figura 5, outras anotações usadas no córpus, em palavras que não são verbos, como, por exemplo: *, que indica que a palavra seguinte designa um participante; &, no início de palavras que devem ser descartadas da computação, por serem imitações ou hesitações; @tag, que refere partículas interrogativas que pedem confirmação no final de um enunciado, como né@tag; e parênteses, que denotam fonemas que foram omitidos na fala, como em p(r)ato.

6.1 Ambigüidades do sistema de verbos do Português

Os casos em que as regras são ambíguas se revelam na resposta do programa. Isso era de se esperar, pois a reprodução pelo programa das ambigüidades do sistema de verbos do português mostra que seu algoritmo corresponde a este sistema. Cabe documentar aqui as ambigüidades do sistema de verbos do português relacionadas aos sufixos modo--temporais.

As ambigüidades cuja resolução é complicada são justamente causadas pela ausência de morfemas específicos que distingam formas verbais. Aliás, quando essas formas ocorrem em um texto, nem sempre é claro para o leitor o tempo verbal em que elas estão. A Figura 5 reproduz a ambigüidade do sistema de verbos no que se refere à ausência de sufixo modo--temporal tanto para 1P-PI como para 1P-PPI. Nesse caso, somente o contexto poderá desambiguar a forma verbal.

Por exemplo, quando a ambigüidade ocorre com o modo imperativo, o contexto pode facilitar a desambiguação ou encarregar-se dela. As formas imperativas afirmativas normalmente ocorrem no início do enunciado ou logo após um vocativo, ao qual sucede uma vírgula, e ocorrem com a segunda pessoa do discurso, isto é, segunda ou terceira pessoas gramaticais. Elas também ocorrem após "por favor" – que não ocorre no córpus de trabalho – e após outras poucas expressões semelhantes. Os morfemas modo-temporais do imperativo negativo coincidem com os do presente do subjuntivo, no entanto, o imperativo negativo, além de estar no mesmo contexto do imperativo afirmativo, sempre vem acompanhado do advérbio de negação "não", de modo que se facilita a resolução da ambigüidade pelo contexto.

A ambigüidade causada pelo uso das flexões de 3S e 3P para as formas "você" e "vocês" é de resolução mais complicada em alguns casos, contudo, nesses casos, a pessoa gramatical normalmente é expressa no enunciado, de maneira que novamente o contexto facilita a desambiguação. Por exemplo, a forma verbal "mostra", do enunciado da linha 940 do córpus:

0940: *MOT: depois você mostra@v p(a)r(a) o papai .

gerava as duas saídas seguintes, das quais nenhuma era correta, pois o pronome subjetivo expresso na sentença não deixa dúvida de que não se trata de imperativo, mas se trata de 2S:

Quadro 2: Resposta inicial do programa à entrada "mostra".

(RAD+VT)		SMT	SNP	Tema&TM&PN
mostr	a	Ø	Ø	mostra&PI&3S
mostr	a	Ø	Ø	mostra&IMA&2S

Observe-se que à forma verbal "mostra" não está agregado morfema modo-temporal nem número-pessoal – ambos são zero. Como diferenciar tempo/modo e pessoa/número, então, se a forma verbal não os expressa? Para resolver esse caso, criou-se uma rotina computacional que verifica o enunciado, à

procura das formas "você", "vocês" e "a_gente".
O pronome "você" ocorre 325 vezes no córpus, e a criança usa-o duas vezes. Por exemplo:

0044 *INV: ah@i (.) você acendeu@v a luz ?
5516 *CHI: vo(u)@va liga(r)@v p(r)a você .

O pronome "vocês" ocorre 14 vezes, e a criança não o usa. Por exemplo:

2855 *ISI: vocês conseguem@va sentar@v os dois juntos ou +...

A forma composta "a gente" somente é pronominal se as duas palavras que a compõem estiverem nessa seqüência e precederem um verbo ou precederem a partícula "se" e/ou um ou mais advérbios antes desse verbo. Para evitar ambigüidade, no córpus, as duas palavras que a compõem aparecem ligadas por "_". No córpus, ela aparece 41 vezes, todas nessa situação. Por exemplo:

1402 *INV: a_gente se diverte@v ,, né ?

O funcionamento dessa rotina computacional consta, em forma de fluxograma, na Figura 6. Depois dessa complementação, a resposta do programa à situação do Quadro 2 é: mostra&PI&2S.
Obter tal distinção nem sempre é fácil, mesmo porque há casos, como visto, em que a pessoa gramatical não é expressa ou está distante do verbo, o que não garante que ela seja seu sujeito. Apesar disso, a grande maioria dos casos fica resolvida com a verificação do contexto do enunciado. Na rotina computacional demonstrada no fluxograma anterior, foi implementada uma instrução para que seja ignorada a partícula "se" anteposta a um verbo, de forma que o verbo "diverte" do enunciado da linha 1402, do exemplo anterior, é corretamente analisado pelo programa: diverte&PI&1P. A análise completa do programa mostra que, nesse caso, SMT e SNP são Ø e que há alomorfia da VT da 3.ªC, que passa de "i" ("divert**ir**") para "e" ("diverte").

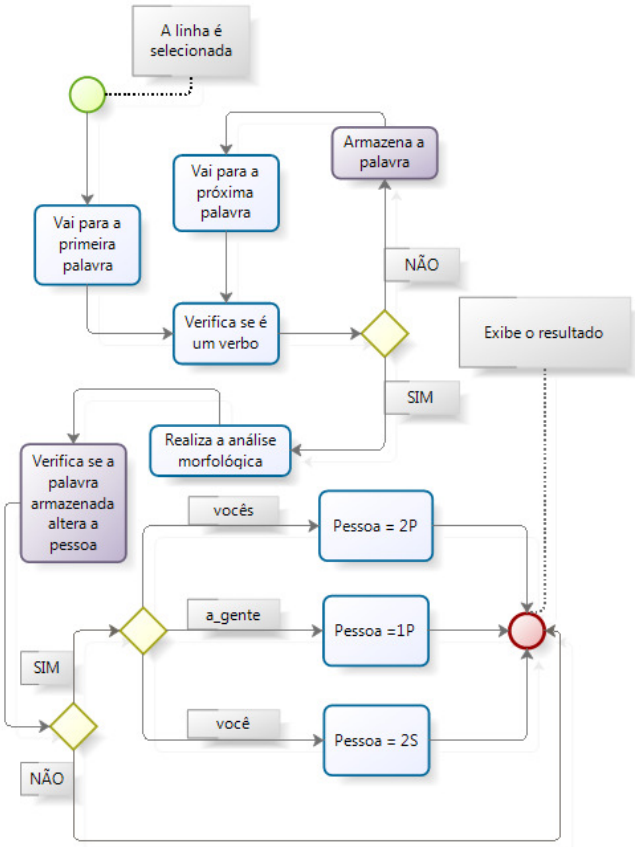

Figura 6: Fluxograma da função que verifica "você", "vocês" e "a_gente" no enunciado.

À medida que a programação avança, revelam-se mais questões a serem tratadas. Assim, a resolução completa das ambigüidades relacionadas às desinências modo-temporais passa pela consideração das desinências número-pessoais e das pessoas gramaticais em si, pois mesmo os tempos verbais não ambíguos dependem das flexões número-pessoais para sua correta e completa análise, bem como do reconhecimento das palavras de outras classes gramaticais que circundam o verbo. Somente a partir disso será possível reduzir ao mínimo ou, talvez, eliminar – a pesquisa dirá – as ambigüidades do sistema de verbos do português do Brasil ocasionadas pelo comportamento dos morfemas modo-temporais.

7 Conclusão e perspectivas

A fase do analisador morfológico automático para verbos do português aqui documentada descreve a automatização das desinências modo-temporais, assim como aborda as ambigüidades provocadas pelo comportamento desses morfemas e apresenta soluções para a

maioria desses casos. No entanto, algumas ambigüidades persistem e, vale dizer, outras podem aparecer. A criação de regras computacionais para verificar o contexto do enunciado foi a principal solução adotada, e o próximo passo será estudar o comportamento dos morfemas número-pessoais, para complementar a desambiguação. Para essa tarefa, já se vislumbram novas regras, que levam em conta, sobretudo e novamente, o contexto do enunciado.

Como se percebe, o trabalho está em evolução constante, de forma que algumas respostas e conclusões somente poderão ser fornecidas ao fim de todas as etapas. Nesse trajeto, pode haver redefinições e redirecionamentos, frutos de aprendizado e testagem empírica.

Agradecimento

Este trabalho é desenvolvido pelo Laboratório de Produtividade Lingüística Emergente da UFSC (LAPLE), com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), entidade do governo brasileiro voltada para a formação de recursos humanos, à qual agradecemos.

Referências

Câmara Jr., Joaquim Mattoso. 1986. *Estrutura da língua portuguesa.* 26.ed. Petrópolis, RJ: Vozes.

Câmara Jr., Joaquim Mattoso. 1976. *História e estrutura da língua portuguesa*. 2.ed. Rio de Janeiro: Padrão.

Childes – Child Language Data Exchange System. 1991-2011. *Clan*: Computerized Language Analysis. http://childes.psy.cmu. edu/clan/

Childes. 2011. Index of Data. *pau003.cha*. http://childes.psy.cmu.edu/data/Romance/ Portuguese/ Florianopolis.zip

Costa, Richard Fernando S. & Scliar-Cabral, Leonor. 2011. Regularização do sistema verbal pela criança. *Anais do Simpósio Internacional Linguagens e Culturas: Homenagem aos 40 anos dos programas de Pós-graduação em Lingüística, Literatura e Inglês da UFSC (SILC)*, Florianópolis, Brasil.

MacWhinney, Brian. 2003-2011. Child Language Data Exchange System. http://childes.psy.cmu.edu/

MacWhinney, B. 2000. The CHILDES Project: Tools for Analyzing Talk. 3rd. ed. Mahwah, NJ: Lawrence Erlbaum Associates. http://childes.psy.cmu.edu/manuals/chat.pdf

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

Scliar-Cabral, Leonor. 2009. *Análise morfológica Automática do Português*. CAPES-UFSC, Florianópolis.

Scliar-Cabral, Leonor. 2008. Codificação da morfologia do PB e análise da fala dirigida à criança. *Fórum Lingüístico*, vol. 5(2), pp.69-82, Florianópolis.

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

Scliar-Cabral, Leonor. 2003. *Princípios do sistema alfabético do português do Brasil*. São Paulo: Contexto.

Scliar-Cabral, Leonor & Vasilévski, Vera. 2011. Descrição do português com auxílio de programa computacional de interface. *Anais da II Jornada de Descrição do Português (JDP)*, Cuiabá, Brasil.

Vasilévski, Vera & Araújo, Márcio J. 2010-2011. *Laça-palavras:* sistema eletrônico para descrição do português brasileiro. LAPLE-UFSC, Florianópolis. https://sites.google.com/ site/sisnhenhem/

Vasilévski, Vera. 2011a. O hífen na separação silábica automática. *Revista do Simpósio de Estudos Lingüísticos e Literários - SELL*, vol. 1(3), pp.657-676, Uberaba.

Vasilévski, Vera. 2011b. Programa para processamento automático das unidades verbais do PB. Análise automática da morfologia do PB (Plataforma CHILDES): aquisição da morfologia verbal. Em *VII*

Congresso Internacional da Abralin, Curitiba, Brasil.

Vasilévski, Vera. 2011c. An automatic system for verb morphological analysis of BP. Em *VIII Encontro Inter-Nacional de Aquisição da Linguagem (ENAL)*, Juiz de Fora, Brasil.

Vasilévski, Vera. 2011d. Diferenças entre Input e Intake: evidências na aquisição de pronomes interrogativos. *Anais do Simpósio Internacional Linguagens e Culturas: Homenagem aos 40 anos dos programas de Pós-graduação em Lingüística, Literatura e Inglês da UFSC (SILC)*, Florianópolis, Brasil.

Vasilévski, Vera. 2010. Divisão silábica automática de texto escrito baseada em princípios fonológicos. *Anais do III Encontro de Pós-graduação em Letras da UFS (ENPOLE)*, São Cristóvão, Sergipe, Brasil.

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

Vasilévski, Vera; Scliar-Cabral, Leonor & Araújo, Márcio J. 2012. Automatic Analysis of Portuguese Verb Morphology: Solving Ambiguities Caused by Thematic Vowel Allomorphs. In *The 10th International Conference on the Computational Processing of Portuguese (PROPOR)*, Coimbra, Portugal, April 17-20.