Novo dicionário de formas flexionadas do UNITEX-PB Avaliação da flexão verbal

Oto A. Vale[1,2], **Jorge Baptista**[3,4]

[1]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – São Carlos – SP – Brasil – 13.565-905

[2]CENTAL – Université Catholique de Louvain (UCL)
Louvain-la-Neuve – Bélgica – B-1348

[3]Faculdade de Ciências Humanas e Sociais – Universidade do Algarve (UAlg)
Campus Gambelas – Faro – Portugal – P-8005-139

[4]Instituto de Engenharia de Sistemas e Computadores (INESC-ID Lisboa/L2F)
Lisboa – Portugal – P-1000-029

otovale@ufscar.br, jbaptis@ualg.pt

Abstract. *This paper describes the new version of the dictionary of inflected forms of Unitex-PB, adapted to the* Acordo Ortográfico *de 1990. Its also presents the evaluation of the verbal forms, which was based in the guidelines established in the first joint evaluation on morphologic analysis of Portuguese (*Primeiras Morfolimpíadas do Português*), held in 2003.*

Resumo. *Neste trabalho descreve-se a nova versão do dicionário de formas flexionadas do Unitex-PB, adaptado ao* Acordo Ortográfico *de 1990. Apresenta ainda a avaliação das formas verbais, que foi realizada a partir dos parâmetros utilizados nas* Primeiras Morfolimpíadas para o Português *(2003).*

1. Introdução

A criação e manutenção de recursos lexicais continua a ser um tema maior no Processamento de Linguagem Natural (PLN). No que diz respeito ao português do Brasil, dentre os diversos recursos criados, o dicionário de formas flexionadas estabelecido por [Muniz et al. 2005] com o sistema UNITEX continua a ser a maior referência de base livremente disponível. Esse recurso foi criado a partir do léxico do REGRA [Nunes et al. 1996, Martins et al. 1998], para os substantivos, adjetivos e advérbios, e a partir da listagem de verbos e de paradigmas de conjugação verbal de [Vale 1990]. Uma revisão recente daquele léxico [Calcia et al. 2014] efetuou sua adaptação ao *Acordo Ortográfico*[1] de 1990, acrescentando também as formas verbais acompanhadas de pronomes pessoais clíticos (em ênclise e mesóclise), que não constavam da versão anterior.

No presente trabalho, procura-se fazer uma avaliação da flexão verbal dessa nova versão do léxico do Unitex. Para tanto, buscou-se um standard utilizado pelas *Primeiras Morfolimpíadas* para o português, organizadas pela Linguateca, de março a junho de 2003 [Santos and Costa 2003]. Na seção seguinte, faz-se uma breve apresentação do Unitex

[1]http://www.portaldalinguaportuguesa.org/acordo.php [2015-08-10]; todos os URL foram validados nesta data.

dos grafos utilizados para criar as formas conjugadas, mostrando as diferenças com a versão anterior. Na seção 3 descreve-se como foi feita a avaliação, sendo os resultados apresentados na seção 4.

2. Descrição do sistema e trabalhos relacionados

O UNITEX [Paumier 2003, Paumier 2014][2] é uma plataforma *open-source* de desenvolvimento de recursos linguísticos, que funciona igualmente como um processador de corpus, baseada em tecnologia de máquinas de estados finitos, e que tem como característica principal a utilização de recursos linguísticos, tais como dicionários e gramáticas locais, bem como uma interface gráfica amigável para o desenho e construção de grafos, que permite gerar de forma automática os respectivos autômatos e transdutores.

Sua versão 3.1 possui recursos para 22 línguas[3]. Os recursos lexicais disponibilizados variam em cobertura e granularidade de língua para língua. Os recursos mais completos são os do francês e do português do Brasil. De fato, para essas línguas pode-se encontrar os dicionários completos de lemas e os grafos ou os transdutores de flexão, que possibilitam a criação do dicionário de formas flexionadas.

Os grafos de flexão disponíveis até a versão 3.1 foram descritos por [Muniz et al. 2005]. Foram então criados 378 grafos para os substantivos, 242 para os adjetivos e mais de 70 para palavras gramaticais como preposições, conjunções, determinantes, numerais e pronomes. Foram utilizados também os 102 modelos de flexão de [Vale 1990] para os verbos. Assim, a partir do dicionário de lemas (DELAS-PB) de 61.335 entradas, gerava-se um total de 878.095 formas flexionadas, que constituiam o DELAF-PB.

No que se refere aos verbos, aquela versão aproveitou as gramáticas de flexão de [Vale 1990], que havia usado a metodologia de [Courtois 1990]. Assim foram criados os transdutores automaticamente, sem passar pelo desenho e construção dos grafos de flexão.

Ao adotar essa solução, deixou-se de incluir as formas enclíticas dos verbos. De fato, [Vale 1990] não havia feito a descrição dessas formas. Essa decisão havia sido tomada por se entender que seria necessário um estudo sintático estabelecendo os verbos suscetíveis de serem afetados por esse fenômeno.

Por outro lado, o tratamento das formas verbais com clíticos está, como é obvio, ligado às opções de cada sistema relativamente ao processo de *tokenização* (ou '*atomização*') dos textos. Como se trata de um passo essencial das fases iniciais do processamento dos textos, várias consequências decorrem das decisões tomadas neste momento, nomeadamente o tratamento das formas com clíticos. Muitos sistemas de PLN adotam o critério geral de reunir num único *token* as formas ligadas por hífen[4]. Ora, por defeito, o UNITEX baseia a tokenização dos textos na lista de carateres do alfabeto da

[2]Disponível em `www.unitexgramlab.org`

[3]São distribuídos com o sistema recursos linguísticos para o alemão, árabe, coreano, espanhol, finlandês, francês, georgiano antigo, grego antigo, grego moderno, inglês, italiano, latim, malgache, norueguês bokmal, norueguês nynorsk, polonês/polaco, português europeu, português do Brasil, russo, sérvio (com alfabeto cirílico e alfabeto latino) e tailandês.

[4]Trata-se, aqui, de uma simplificação um pouco excessiva, é verdade, já que o processo de tokenização pode ser modelado de diversas formas, dependendo do sistema, e algumas delas podem implicar uma elevada granularidade na decisão sobre as formas a reunir num único token.

língua de trabalho, considerando todos os restantes como separadores (além dos dígitos que têm um tratamento à parte). A fim de considerar as formas verbais com clíticos como um único *token*, é possível, no entanto, adaptar a lista do alfabeto da língua de trabalho, acrescentando-lhe o hífen. Sem querer aqui entrar na discussão sobre os méritos e inconvenientes de cada uma das opções, foi essa a solução adotada nesta nova versão do dicionário.

Entretanto, [Ranchhod et al. 1999] ao apontarem para o português europeu algumas dificuldades no estabelecimento da listagem dos verbos que poderiam ser conjugados com as formas enclíticas, salientavam que a descrição da morfologia dessas formas deveria ser feita antes dessa descrição sintática. Idêntica solução foi adotada para o sistema STRING [Mamede et al. 2012] por [Vicente 2013]. Em outras palavras, considera-se que a descrição da flexão verbal (e da sua variação em função das combinatórias com pronomes pessoais clíticos) é um problema que deve ser primeiro resolvido a um nível estritamente morfológico, sendo depois o nível sintático responsável pelas restrições combinatórias que resultam das diferentes construções em que o verbo pode entrar (ou, dito de outro modo, das diferentes valências que o verbo apresentar) [5].

[Calcia et al. 2014] realizaram uma adaptação dos grafos de flexão e da listagem do DELAS-PB para a nova ortografia, resultante do *Acordo Ortográfico* de 1990, além de terem procedido à atualização do dicionário. Das 878.095 formas, 1.287 sofreram algum tipo de modificação. Além disso, foram introduzidas 7.900 novas entradas[6]. Naquele trabalho, apresentou-se uma primeira versão dos grafos de conjugação dos verbos do português do Brasil com as formas enclíticas e mesoclíticas; dito de outra forma, cada paradigma verbal foi descrito com as formas enclíticas e mesoclíticas. Mais concretamente, cada grafo de conjugação verbal foi construído com o auxílio de subgrafos, que descreviam as particularidades de cada tempo verbal e introduziam também os clíticos associados a cada forma, como se pode ver nas Figuras 1 e 2.

O grafo da Fig. 1 representa o paradigma de flexão de verbos regulares de tema em *-a*, como *cortar* e interpreta-se do seguinte modo: o operador `L` (do ing. *left*) indica o número de carateres a retirar ao final do lema; as alterações à terminação da palavra aparecem nas caixas e sob estas os valores gramaticais correspondentes; assim, a partir de um lema como *cortar*, a primeira linha, no topo do grafo, produz a forma do gerúndio (`G`) *cortando*, que corresponde à remoção (`L`) do *-r* do final do lema e a adição da terminação *-ndo*; os códigos convencionais para os valores gramaticais de tempo-modo, pessoa e número foram descritos em [Calcia et al. 2014]. Os tempos-modos verbais associados a cada paradigma de flexão são descritos por meios de subgrafos (caixas cinzentas), de que se apresenta como exemplo, na Fig. 2, o caso do Presente do Indicativo (`P`) dos verbos regulares da 1ª e da 2ª conjugação.

Nesta Figura, à esquerda, pode ver-se, a par de cada flexão em pessoa número, os diferentes conjuntos de pronomes clíticos que se podem combinar com a forma considerada e que se representa por meio de subgrafos (caixas cinzentas; os nomes dos grafos

[5]Uma questão a ser tratada posteriormente diz respeito à adequação da notação dos pronomes das formas enclíticas e mesoclíticas.

[6]Esta nova versão do dicionário e dos respetivos grafos de flexão já estão sendo distribuídos com o UNITEX 3.1 e pode também ser encontrada na página: `http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html`

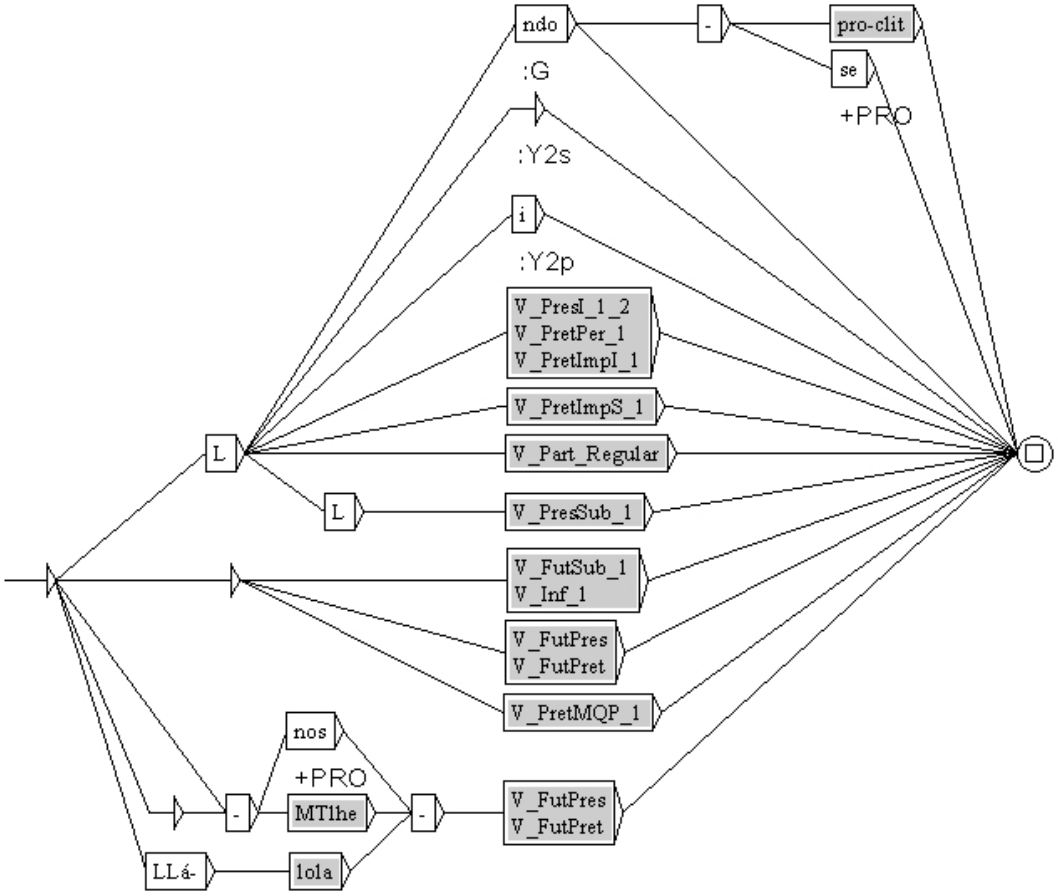

Figura 1. Grafo de flexão `V005.grf` aplicado as verbos regulares da primeira conjugação.

estão a vermelho). Os três grafos auxiliares encontram-se à direita da figura, na linha de cima. Estes apelam, por sua vez a outros grafos, representados abaixo, à direita.

Dada a natureza modular destes fenômenos, a representação por autômatos de estados finitos permite, assim, tratá-los de forma bastante econômica e precisa.

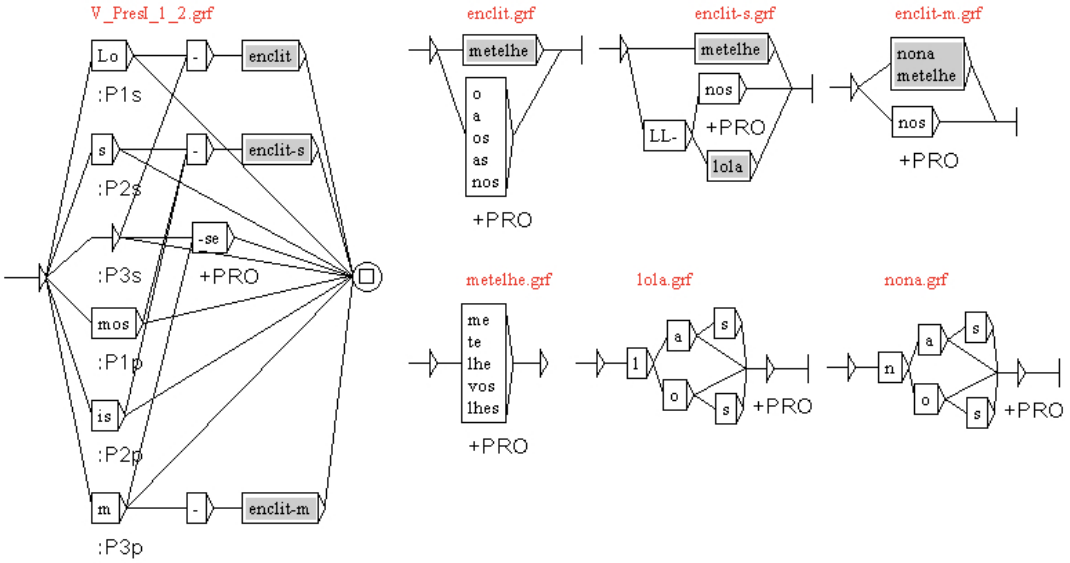

Figura 2. Subgrafo `V_PresI_1_2.grf` de flexão do presente do indicativo dos verbos regulares da primeira e segunda conjugações, com as associações dos clíticos para cada forma.

Um dos problemas apontados em [Calcia et al. 2014] para essa solução de introduzir as formas enclíticas e mesoclíticas como entradas do dicionário das formas flexionadas é a "explosão" do número de formas verbais geradas. De fato, enquanto o DELAF-PB de [Muniz et al. 2005] continha 878.095 formas, a versão inicial do dicionário de [Calcia et al. 2014], produzido a partir das regras que descrevem as formas verbais com clíticos, é de 10.954.724 entradas (7.632.498 formas diferentes), das quais 10.772.850 são formas verbais (7.477.680 formas diferentes). Entretanto, graças ao desenvolvimento dos algoritmos de compressão do UNITEX 3.1, o arquivo `.bin` dessa nova versão do dicionário ocupa agora apenas 778 KB (mais 480 KB do arquivo `.inf` que descreve os códigos associados à compressão), enquanto a versão anterior ocupava 819 KB (mais 208KB do arquivo `.inf`).

3. Avaliação

Para avaliação, utilizaram-se os recursos produzidos para as *Primeiras Morfolimpíadas para o português*, organizadas pela Linguateca em 2003, nomeadamente as formas Lista Dourada[7], que estavam anotadas como verbos (510 linhas) e que serviram de referência para esta campanha de avaliação conjunta. Para uma comparação da saída do DELAF-PB com a Lista Dourada, esta última foi convertida no formato DELA, tendo-se, nomeadamente:

[7] http://www.linguateca.pt/aval_conjunta/morfolimpiadas/ListaDourada.txt

(i) substituído os códigos dos tempo e modos verbais pelos códigos do DELAF-PB;

(ii) substituído as maiúsculas iniciais na Lista Dourada (*e.g. Apoiemos*) por minúsculas, já que as formas das entradas no formato DELA são sempre grafadas em minúsculas;

(iii) desdobrado as 15 formas apresentando dupla grafia na Lista Dourada; trata-se dos casos seguintes:

(a) formas com consoante muda (*e.g. conetar/conectar*);

(b) forma com trema (*e.g. freqüentar/frequentar*, *seqüenciais/sequenciais*); e

(c) as variantes *registar/registrar*;

(iv) substituído o código `V+CL` por `V+PRO` nas 22 formas verbais com clíticos e remoção do clítico do campo do lema;

(v) retirado, ainda, as anotações de uso (‘`raro`’; ‘`lus`’, ‘`bras`’, ‘`afr`’), ou morfológicas, nomeadamente para as formas derivadas e analisadas como tal (‘`deriv`’, ‘`pref`’), e os prefixos envolvidos nas formas derivadas (‘`a`’, ‘`des`’, ‘`in`’, ‘`re`’, e ‘`sub`’).

Note-se, em relação a este último aspecto, que o DELAF-PB não analisa as formas derivadas, limitando-se a registrar essas formas como lemas diferentes no DELAS-PB e a gerar as correspondentes formas flexionadas. Não faria, assim, qualquer sentido tentar avaliar o que o sistema não se propõe a fazer. As 56 formas derivadas foram, portanto, removidas da lista dourada numa segunda fase de avaliação.

Obteve-se, assim, um total de 296 formas que foram analisadas pelo UNITEX com o novo dicionário DELAF-PB.

Da comparação entre a saída do dicionário e a Lista Dourada (referência), é possível obter os seguintes resultados:

Corretos : a saída do dicionário é igual à referência;
Errados : a saída do dicionário é diferente da referência;
Lacunas : a forma e a sua análise na Lista Dourada não estão na saída do dicionário; e
Espúrios : a forma e a sua análise são produzidas pelo dicionário mas não estão na referência.

Para a avaliação consideraram-se as seguintes medidas standard:

Precisão : total de formas corretamente analisadas:
corretos / (corretos + errados + espúrios);
Abrangência (em ing. *recall*) : total de formas corretamente analisadas de entre todas as formas analisadas na Lista Dourada:
corretos / (corretos + lacunas);
Acurácia (do ing. *accuracy*) : total de formas corretamente analisadas:
corretos / (corretos + errados + lacunas);
Medida F média harmônica entre a Precisão e a Abrangência:
*2 * Precisão * Abrangência / (Precisão + Abrangência)*

Tabela 1. Resultados da avaliação[8]

Aval	Cor	Err	Lac	Esp	Prec	Abr	Acur	med-F
A	416	22	85	117	0,795	0,780	0,749	0,787
B	438	22	40	40 (72VN)	0,876	0,796	0,766	0,834

Os resultados "em bruto" são apresentados na Tabela 1, linha (A). Há, no entanto, que considerar alguns aspectos que alteram consideravelmente a interpretação destes resultados:

(i) as entradas cuja forma ou lema não corresponde à norma ortográfica brasileira devem ser consideradas como *verdadeiros-negativos*, já que o dicionário não se propõe a descrevê-las. Nesses casos incluem-se, por exemplo, as formas com consoantes surdas:
`atuais,actuais.V:P2p`,
`conectar,conetar.V:U1s:U3s:W:W1s:W3s`,
`objecto,objectar.  V:P1s`);

(ii) os *lusismos* das formas graficamente acentuadas da primeira pessoa do plural do pretérito imperfeito:
`abandonámos,abandonar.V:J1p`;

(iii) a variante lusa: `registro,registar.V:P1s`;

(iv) as formas com trema na anterior ortografia brasileira: *freqüentar/frequentar*;

(v) a forma *dêem*, cuja ortografia foi igualmente alterada;

(vi) as formas derivadas, isto é, que resultam de uma análise morfológica, e para as quais, na Lista Dourada, se indica como lema a forma de base; estas formas, como já dissemos, deviam ser reconhecidas mas não analisadas pelo DELAF-PB, que não foi concebido para esse fim;

(vii) as formas verbais simples que fazem parte de formas verbais com clítico (*v.g. capacite* em *capacite-se*) e que, pela sua duplicação na saída do dicionário, enviesam os resultados; estas formas não deveriam ser consideradas espúrias, pelo que foram assim ignoradas; saliente-se, contudo, a forma *ir-se-ia*, que recebe duas segmentações pelo sistema (*ir-se* e *ir-se-ia*), pelo que as duas análises associadas ao infinitivo devem continuar a ser tratadas como espúrias;

(viii) finalmente, as 22 formas que foram bem analisadas quanto à categoria e à flexão mas a que o dicionário não foi atribuiu um lema (*v.g. apregoar*, *desmobilizar*, *proceder*, *vagar* e *zoar*); em rigor, trata-se de uma resposta parcialmente correta mas incompleta, pelo que os mantivemos como falsos-positivos.

[8] Aval=avaliação, A:resultados em bruto, B:resultados corrigidos; Cor=corretos, Err=errados, Lac=lacunas, Esp=espúrios; Prec=precisão, Abr=abrangência, Acur=acurácia, Med-F=medida-F.

Numa análise mais fina destes resultados, verificamos ainda alguns aspectos que merecem um tratamento diferenciado.

Em alguns casos, os erros resultam de incompletude da Lista Dourada. Assim, por exemplo, alguns lemas raros não tinham sido incluídos na referência, *v.g. iriar*, *presar*, *rer*, *revir*, *valar* e *vivar* .

Por vezes, essas lacunas são flexões exclusivas da variante brasileira, *v.g. pega* e *pegas*, como particípios passados de *pegar*, por derivação regressiva.

Outros casos, aparentemente espúrios, resultam de opções linguísticas do DELAF-PB, que sistematicamente diferem das opções da Lista Dourada. Efetivamente, seguindo a tradição gramatical brasileira, o dicionário considera uma flexão do imperativo da terceira pessoa do singular (`Y3s`, *aceite*) e do plural (`Y3p`, *peçam*), que correspondem à forma de tratamento por *você*, e ainda uma flexão de primeira do plural (`Y1p`, *sigamos*).

Pelo contrário, algumas lacunas da referência são perfeitamente assistemáticas, *v.g. ante*, de *antar* (raro), só tem a primeira pessoa do singular do presente do conjuntivo/subjuntivo (`S1s`), mas não a terceira (`S3s`). Em rigor, essas análises espúrias do DELAF-BP são, de fato, ou omissões da Lista Dourada ou, como no caso dos imperativos, decisões do dicionário, conformes à tradição gramatical brasileira, pelo que foram, num segundo momento, tratadas como verdadeiros-negativos.

Do lado das lacunas do dicionário, este exercício permitiu detectar algumas inconsistências, que foram posteriormente corrigidas. Certos lemas, alguns raros não estavam no dicionário, *v.g. devir*, *frequentar*, *incendiar*, *injustiçar*, *negociar*, *parir*, *redar*, *redobrar*, *reinterpretar*, *reversar*, *subdesenvolver*, *surfar*, *surpresar* e *travestir*. Note-se que alguns destes casos são formas derivadas regularmente.

Certas flexões irregulares também não foram incluídas no dicionário, como é o caso dos particípios *expulsas*, de *expelir*, e *junto*, de *juntar*.

Note-se que os casos de verdadeiros-negativos não fazem parte dos quatro tipos de resultados considerados na primeira fase da análise. Assim, estes casos deverão ser acrescentados ao denominador da abrangência e da acurácia. Os resultados corrigidos estão também apresentados na Tabela 1, linha (B).

Não sendo as condições neste momento exatamente as mesmas das que tiveram lugar na campanha de avaliação das Morfolimpíadas, cabe, no entanto, aqui uma breve comparação entre os resultados deste exercício com alguns dos resultados daquela campanha. Assim, usando precisamente as mesmas medidas de avaliação das Morfolimpíadas, nomeadamente as que descrevem os resultados que correspondem à Tabela 1 ("Comparação com a lista dourada total, sem lema nem outro") da página dos *Resultados*[9], é possível chegar aos seguintes resultados de avaliação do desempenho do dicionário `delaf-pb` na análise das formas verbais da Lista Dourada, e que se apresentam na Tabela 2.

4. Conclusão

Tendo em vista os resultados apresentados nesta avaliação, nota-se que o desempenho dessa nova versão do DELAF-PB na análise das formas verbais é bastante satisfatório em

[9] `http://www.linguateca.pt/aval_conjunta/morfolimpiadas/comp_dourada_fig.html`

Tabela 2. Avaliação do desempenho do DELAF-PB na análise das formas verbais da Lista Dourada usando as medidas das *Morfolimpíadas*

Avaliação	Relativa	Absoluta
Formas Comparadas	270	286
Análises na Lista Dourada	523	555
Análises no DELAF-PB	523	
Análises comuns	438	
Precisão	0,837	
Cobertura	0,837	0,789

relação aos desafios propostos nas *Morfolimpíadas*. A precisão está dentro dos parâmetros obtidos pelos restantes sistemas (para o conjunto de todas as categorias), e a cobertura ficou acima dos valores médios que tinham sido ali alcançados.

Cabe também notar que essa avaliação permitiu perceber algumas lacunas e inconsistências presentes no DELAF-PB. Sem querer fazer uma lista exaustiva, pode-se exemplificar com a introdução indevida do pronome reflexivo da terceira pessoa *–se* em formas de primeira e segunda pessoa, gerando formas claramente incorretas. Outro exemplo, desta vez de omissão, foi o fato de a introdução das formas enclíticas, em alguns tempos, ter impedido a geração de formas corretas, como, por exemplo, as terceiras pessoas do singular do mais-que-perfeito do indicativo em praticamente todos os verbos regulares.

Esses resultados permitem apontar para um aperfeiçoamento do dicionário para uma próxima versão, a ser distribuída brevemente com o sistema UNITEX.

Agradecimentos

Este trabalho foi parcialmente financiado pela FAPESP, pela CAPES e pelo CNPq (Brasil), pela Fundação para a Ciência e a Tecnologia (ref. UID/CEC/50021/2013, Portugal), e pelo Dicionário Informal (www.dicionarioinformal.com.br)

Referências

Calcia, N. P., Kucinskas, A. B., Muniz, M., Nunes, M. G. V., and Vale, O. A. (2014). Révision et adaptation des dictionnaires et graphes de flexion d'Unitex-PB à la nouvelle orthographie du portugais. In *3rd UNITEX/GramLab Workshop*, Université de Tours. 3rd UNITEX/GramLab Workshop.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, (87):11–22.

Mamede, N., Baptista, J., and Diniz, C. (2012). String - an hybrid statistical and rule-based natural language processing chain for portuguese. In Demos, P. ., editor, *PROPOR 2012*, Coimbra, Portugal. PROPOR, PROPOR.

Martins, R. T., Hasegawa, R., Nunes, M. G. V., G. Montilha, G., and Oliveira, O. N. (1998). Linguistic issues in the development of REGRA: a grammar checker for Brazilian Portuguese. *Natural Language Engineering*, 4(4):287—307.

Muniz, M. C. M., Nunes, M. G. V., and Laporte, E. (2005). UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. In *Workshop on Technology on Information and Human Language (TIL 2005)*, pages 2059–2068, São Leopoldo, Brazil. SBC.

Nunes, M. G. V., Vieira, F. M. C., Zavaglia, C., Sossolote, C. R. C., and Hernandez, J. (1996). A construção de um léxico de português do Brasil: Lições aprendidas e perspectivas. In *Anais do II Workshop de*

Processamento Computacional de Português Escrito e Falado (PROPOR'96), pages 61—70, CEFET-PR, Curitiba.

Paumier, S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse de doctorat, Université de Marne-la-Vallée, Paris.

Paumier, S. (2014). *Unitex 3.1 - User Manual*. Université de Paris-Est/Marne-la-Vallée - Institut Gaspard Monge, Noisy-Champs.

Ranchhod, E., Mota, C., and Baptista, J. (1999). A computational lexicon of Portuguese for automatic text parsing. In *SIGLEX'99: Standardizing Lexical Resources*, pages 74–80, Maryland, USA. SIGLEX/ACL: Special Interest Group on the Lexicon of the Association for Computational Linguistics and the National Science Foundation, ACL/SIGLEX.

Santos, D. and Costa, L. (2003). Morfolimpíadas - apresentação detalhada da metodologia e dos problemas identificados. In *AvalON'2003*, Faro. Linguateca/Universidade do Algarve.

Vale, O. A. (1990). Dictionnaire électronique des conjugaisons des verbes du portugais du Brésil. Rapport Technique 27, LADL-Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7, Jussieu, Paris.

Vicente, A. M. F. (2013). Lexman: um segmentador e analisador morfológico com transdutores. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.