Resumo

Este relatório apresenta a ferramenta de análise morfossintática `anali`, a qual foi desenvolvida no NILC (Núcleo Interinstitucional de Lingüística Computacional) como resultado da união de outras duas ferramentas de Processamento de Língua Natural: o etiquetador MXPOST e a ferramenta de análise de corpus Unitex. Nesse sentido, `anali` representa um ganho em relação ao que é produzido pelas ferramentas citadas, em dois sentidos. Por um lado, enriquece a saída de MXPOST inserindo mais informação a respeito da análise retornada por esse etiquetador; e, por outro, desambigua a saída de Unitex ao definir qual das várias análises existentes em seus dicionários eletrônicos é a melhor, em cada caso. Além disso, `anali` pode operar em três modos distintos: etiquetação (com base apenas na saída de MXPOST), análise morfossintática (com base apenas na saída de Unitex) ou ambos.

Sumário

1 Introdução

Este relatório apresenta a ferramenta de análise morfossintática `anali` desenvolvida no NILC (Núcleo Interinstitucional de Lingüística Computacional), no âmbito do projeto ReTraTos[1], como o resultado da união de duas outras ferramentas de Processamento de Língua Natural: o etiquetador MXPOST (RATNAPARKHI, 1996) e a ferramenta de processamento de corpus Unitex (PAUMIER, 2006).

O projeto ReTraTos visa a indução automática de conhecimento lingüístico útil para a tradução automática do português do Brasil (`pt`) e outros dois idiomas – espanhol (`es`) e inglês (`en`) – agrupados em dois pares de tradução: `pt`–`es` e `pt`–`en`. Tais recursos, representados na forma de regras de tradução e léxicos bilíngües, são induzidos automaticamente a partir de textos paralelos etiquetados morfossintaticamente e alinhados sentencial e lexicalmente. Para tanto, ferramentas automáticas são utilizadas para preparar os textos paralelos a serem utilizados por ReTraTos, são elas: (1) o alinhador sentencial `TCAalign`, (2) o alinhador lexical `LIHLA`[2] e (3) ferramentas de análise morfossintática diferentes para os corpora paralelo `pt`–`es` (módulos presentes no tradutor automático `Apertium`[3]) e `pt`–`en` (ferramenta apresentada aqui, `anali`).

Como mencionado anteriormente, `anali` foi gerada como o resultado da união de duas outras ferramentas: MXPOST e Unitex. O etiquetador MXPOST foi

[1] O projeto ReTraTos tem o apoio de FAPESP, CAPES e CNPq.

[2] Informações a respeito dos alinhadores sentencial e lexical usados no projeto ReTraTos podem ser obtidas em `http://www.nilc.icmc.usp.br/nilc/projects/aligners.htm`.

[3] `www.apertium.org`

desenvolvido por Ratnaparkhi (1996) e utiliza o modelo de Máxima Entropia para determinar qual etiqueta de *Part-Of-Speech* (POS) cada *token* que compõe um texto de entrada deve receber. Esse modelo probabilístico utiliza os pesos de cada valor de atributo relevante para estimar a probabilidade P(etiqueta|atributos). Na fase de etiquetagem, o algoritmo de busca *beam search* é utilizado para encontrar a seqüência de etiquetas mais provável para todo o período[4].

O etiquetador MXPOST foi escolhido, entre outros etiquetadores disponíveis na Web – como o TreeTagger (SCHMID, 1994) e o etiquetador de BRILL (BRILL, 1995) –, por ter sido o de melhor desempenho em experimentos realizados, no NILC, com o português do Brasil: de 94,39% a 96,98% de precisão nos experimentos realizados com o corpus MAC-MORPHO.[5] Para o idioma inglês, MXPOST apresentou 96,6% de precisão em experimentos realizados com o corpus do *Wall Street Journal* (RATNAPARKHI, 1996).

Tanto nos experimentos com o português quanto nos experimentos com o inglês, o tamanho do corpus de treinamento foi próximo a 1 milhão de palavras: 962.687 palavras no inglês e 977.161 palavras no português. É importante citar que, embora o MXPOST tenha sido escolhido para ser utilizado na construção de `anali`, outro etiquetador também poderia ser utilizado em seu lugar com as devidas alterações no código e parâmetros de entrada.

O Unitex, por sua vez, é uma coleção de recursos e ferramentas lingüísticas (dicionários eletrônicos, gramáticas, etc.) usados para a análise de textos em linguagem natural. Os dicionários eletrônicos do Unitex especificam as palavras simples e compostas de uma língua juntamente com seus lemas e um conjunto de códigos gramaticais (semânticos e flexionais). Esses dicionários são representados no formalismo DELA e estão disponíveis[6] para vários idiomas entre eles Inglês,

[4]Um período é composto por uma ou mais orações e/ou frases e tem como características: (1) apresentação de um sentido ou significado completo e (2) encerrar-se por meio de certos símbolos de pontuação (AIRES, 2000).

[5]Os detalhes dos experimentos realizados com MXPOST para o português do Brasil podem ser consultados em: `http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm`.

[6]A ferramenta de processamento de corpus Unitex, bem como os dicionários eletrônicos no

Espanhol e Português[7] (PAUMIER, 2006).

Embora a precisão obtida por MXPOST na etiquetação de textos em português e inglês seja considerada boa, a motivação para a criação de uma nova ferramenta surgiu da análise do conjunto de etiquetas utilizado para o português (ALUÍSIO et al., 2003) e o inglês (Penn tagset, (MARCUS et al., 1993)). A partir dessa análise constatou-se que ambos os conjuntos de etiquetas não trazem todas as informações morfossintáticas de interesse para o projeto ReTraTos. Assim, optou-se por utilizar os dicionários eletrônicos de Unitex para "enriquecer" as etiquetas de MXPOST e, até mesmo, acrescentar novas análises às palavras.

Além do Unitex e do MXPOST, opcionalmente, regras de desambiguação também podem ser utilizadas para auxiliar a seleção da melhor análise a ser atribuída a uma dada palavra. Tais regras, especificadas em um arquivo texto seguindo formatos pré-definidos (veja seção 2.3), são passadas como parâmetro de entrada para o programa e são processadas, em tempo de execução, para desambiguar categorias específicas.

Nesse sentido, este relatório descreve, no capítulo 2, a ferramenta `anali` desenvolvida com o intuito de unir as análises morfossintáticas de Unitex com as etiquetas de MXPOST. Esse processo de "enriquecimento" da saída de MXPOST (ou desambiguação das análises de Unitex) pode ser aplicado em textos escritos em qualquer idioma para o qual estejam disponíveis dicionários de Unitex e modelos treinados de MXPOST. Por fim, o capítulo 3 traz algumas considerações finais.

O Apêndice A apresenta os arquivos de mapeamento definidos para o inglês (seção A.1) e o português (seção A.2). O arquivo de mapeamento é utilizado para mapear as etiquetas de MXPOST nas etiquetas de Unitex e permitir, assim, que a intersecção entre as saídas dessas duas ferramentas seja possível.

O Apêndice B, por sua vez, traz os arquivos com os conjuntos de regras de

formalismo DELA, podem ser obtidos em `http://www-igm.univ-mlv.fr/~unitex`.

[7] A construção dos dicionários eletrônicos no formato do Unitex para o Português do Brasil foi resultado de um projeto de mestrado desenvolvido no NILC (MUNIZ, 2004).

desambiguação definidos para o inglês (seção B.1) e o português (seção B.2). Nesses arquivos, cada regra de desambiguação deve seguir um dos três formatos pré-definidos apresentados na seção 2.3 do capítulo 2.

Por fim, o Anexo A apresenta os conjuntos de etiquetas utilizadas por MX-POST para o inglês (seção A.1) e o português (seção A.2); enquanto o Anexo B apresenta os conjuntos de etiquetas utilizados por Unitex para ambos os idiomas (seção B.1 para o inglês e seção B.2 para o português).

2 A ferramenta de análise morfossintática `anali`

Como mencionando anteriormente, a ferramenta `anali` foi desenvolvida para unir as funcionalidades de MXPOST e Unitex e, portanto, pode ser operada em 3 modos: (1) etiquetação (com base apenas na saída de MXPOST), (2) análise morfossintática (com base apenas na saída de Unitex) ou (3) ambos.

No primeiro modo (etiquetação), `anali` faz uma chamada ao MXPOST para que este atribua, a cada palavra (e alguns símbolos de pontuação), a melhor etiqueta morfossintática de acordo com sua estratégia de etiquetação. No segundo modo (análise), `anali` utiliza ferramentas e recursos de Unitex para gerar os dicionários das palavras simples que ocorrem no texto de entrada e, para cada palavra nesse texto, atribui a seqüência de possíveis análises morfossintáticas. Por fim, no terceiro modo (etiquetação e análise), os dois processos citados anteriormente são realizados e o resultado é a interseccção de suas saídas.

A escolha do modo de operação de `anali` é feita pelo usuário por meio de um dos 5 parâmetros de entrada possíveis:

1. `-m [e|a|ae]` (obrigatório) – modo de operação de `anali`: etiquetação (`-m e`), análise (`-m a`) e ambos (`-m ae`);

2. `-a <caminho_arquivo>` (obrigatório) – caminho completo do arquivo a ser processado;

3. `-d <diretório>` (obrigatório) – diretório com os dados específicos do idioma sendo processado, ou seja, o diretório deve conter o modelo de MXPOST (em um subdiretório denominado "mxpost") e os dicionários eletrônicos de Unitex (em um subdiretório denominado "unitex");

4. `-map <caminho_arquivo>` (obrigatório apenas para o modo `ae`) – caminho completo do arquivo com o mapeamento das etiquetas de MXPOST para Unitex;

5. `-des <caminho_arquivo>` (opcional) – caminho completo do arquivo com as regras de desambiguação;

6. `-t` (opcional) – indicação de que todas as opções de análise possíveis, em um determinado caso, devem ser impressas. Isso ocorre quando a intersecção entre MXPOST e Unitex não tem sucesso (não há nenhuma análise em comum) ou essa intersecção não resulta em uma única opção de análise.

O formato do arquivo de saída de `anali` varia de acordo com o modo de processamento selecionado, como apresentado na Tabela 1.

Tabela 1: Formato do arquivo de saída de acordo com o modo de processamento

Modo	**Formato**	**Exemplos**
`e`	palavra_ETIM	team_NN equipe_N
`a`	palavra{A_1}{A_2}...{A_n} A_i = lema.ETIU:atr_1:...:atr_n atr_i = seqüência de caracteres em que cada caractere representa um atributo	team{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} equipe{.N:fs}{equipar.V:S1s:S3s:Y3s}
`ae`	palavra/A_1 palavra/{A_1}{A_2}...{A_n}	that/.PREP that/{.PREP}{.CONJS}{.CONJ}{.ADV} {.DET+Ddem:s}{.PRO+Pdem:s} pela/{pelo.PREPXDET+Art+Def:fs} {pelo.PREPXPRO+Dem:fs}

No modo de etiquetação (`e`), cada *token* do arquivo de entrada é processado e a ele é atribuída uma etiqueta do conjunto de etiquetas de MXPOST (ETIM)

precedida pelo caractere "_".[1] No modo de análise (`a`), cada palavra é seguida de uma ou mais análises morfossintáticas possíveis (A_i) delimitadas pelos caracteres "{" e "}". Cada análise A_i é formada pelo lema da palavra (se este for diferente da forma superficial), seguido pelo caractere ".", uma etiqueta do conjunto de etiquetas de Unitex (ETIU) e, opcionalmente, o caractere ":" como separador dos conjuntos de atributos (atr).[2] No modo que combina a funcionalidade dos dois modos anteriores (`ae`), cada *token* é seguido pelo caractere "/" e uma análise (A_1), se a opção `-t` não foi especificada; ou todas as análises possíveis (delimitadas pelos caracteres "{" e "}"), se a opção `-t` foi especificada.

A seguir são apresentadas as tarefas desempenhadas em cada modo de processamento: etiquetação (seção 2.1), análise (seção 2.2) e ambos (seção 2.3).

2.1 Etiquetação

Para realizar a etiquetação, `anali` simplesmente faz uma chamada ao etiquetador MXPOST passando como parâmetro o conteúdo do arquivo de entrada sem as etiquetas de início e fim de sentenças, se estas estiverem presentes. Ao final do processo de etiquetação, os delimitadores de sentenças são introduzidos, no arquivo de saída, em seus devidos lugares para a geração do arquivo de saída.

Além da simples chamada ao MXPOST, nesse modo também são realizadas duas tarefas que diferenciam seu processamento de uma execução normal do MXPOST: uma tarefa de pré-processamento e outra de pós-processamento.

A tarefa de pré-processamento é responsável por separar os *tokens* do arquivo de entrada inserindo espaços em lugares pré-estabelecidos como, por exemplo, antes e depois de caracteres de pontuação mas não entre uma vírgula e os números que a cercam (como no número real 1,5). Esse processo de tokenização diminui as chances do etiquetador considerar erroneamente os *tokens* a serem processados.

[1] Veja o conjunto de etiquetas utilizado por MXPOST no Anexo A.
[2] Veja o conjunto de etiquetas utilizado por Unitex no Anexo B.

A tarefa de pós-processamento, por sua vez, corrige erros do etiquetador relacionados à atribuição de uma etiqueta de palavra a um caractere de pontuação como, por exemplo, no trecho "(_NNP São_NNP Paulo_NNP)_NNP" que é convertido para "(São_NNP Paulo_NNP)". O conjunto de etiquetas utilizado por MXPOST pode ser consultado no Anexo A.

Por padrão, o arquivo de saída gerado no modo `e` possui o mesmo nome do arquivo de entrada, porém com a extensão ".e". Assim, a Tabela 2 apresenta um trecho de um texto em inglês (ingles.txt) etiquetado por meio da execução do comando:

```
> perl anali.pl -m e -a ingles.txt -d Ingles
```

Tabela 2: Três primeiras sentenças de um texto em inglês etiquetado por `anali`

Re_NNP -_: inventing_VBG Benjamin_NNP Franklin_NNP

A_DT team_NN of_IN Brazilian_JJ researchers_NNS produced_VBD artificial_JJ lightning_NN for_IN the_DT first_JJ time_NN in_IN the_DT Southern_NNP Hemisphere_NNP ._.
In_IN the_DT final_JJ days_NNS of_IN November_NNP ,_, the_DT people_NNS at_IN the_DT International_NNP Center_NNP of_IN Research_NNP and_CC Tests_VBZ on_IN Lightning_NNP ,_, of_IN the_DT National_NNP Institute_NNP of_IN Spatial_NNP Research_NNP (Inpe_NNP) ,_, repeated_VBD the_DT experiment_NN done_VBN some_DT 248_CD years_NNS ago_RB when_WRB Benjamin_NNP Franklin_NNP flew_VBD a_DT kite_NN with_IN a_DT spanner_NN on_IN the_DT end_NN of_IN it_PRP to_TO prove_VB that_IN the_DT atmosphere_NN was_VBD full_JJ of_IN electricity_NN ._.
...

A Tabela 3 apresenta um trecho de um texto em português (portugues.txt) etiquetado por meio da execução do comando:

```
> perl anali.pl -m e -a portugues.txt -d Portugues
```

Tabela 3: Três primeiras sentenças de um texto em português etiquetado por `anali`

Reinventando_NPROP Benjamin_NPROP Franklin_NPROP
Uma_ART equipe_N de_PREP pesquisadores_N brasileiros_ADJ produziu_V raios_N artificialmente_ADV pela_NPROP primeira_ADJ vez_PREP no_NPROP Hemisfério_NPROP Sul_NPROP ._.
Nos_PROPESS últimos_ADJ dias_N de_PREP novembro_N ,_, o_ART pessoal_N do_NPROP Centro_NPROP Internacional_NPROP de_NPROP Pesquisa_NPROP e_KC Testes_NPROP de_NPROP Raios_NPROP ,_, do_NPROP Instituto_NPROP Nacional_NPROP de_NPROP Pesquisas_NPROP Espaciais_NPROP (_(Inpe_NPROP)_) ,_, repetiu_V a_ART experiência_N feita_PCP há_PREP 248_NUM anos_N ,_, quando_ADV-KS-REL Benjamin_NPROP Franklin_NPROP empinou_V uma_ART pipa_N com_PREP uma_ART chave_N na_N extremidade_N para_PREP provar_V que_KS a_ART atmosfera_N estava_V carregada_PCP de_PREP eletricidade_N ._.
...

2.2 Análise

A análise morfossintática é produzida com base nos dicionários eletrônicos de Unitex e pode ser dividida em duas etapas: (1) geração do dicionário de palavras simples que ocorrem no arquivo de entrada com base nos dicionários eletrônicos de Unitex e (2) atribuição de todas as análises possíveis a cada palavra do arquivo de entrada.

Para a realização da primeira etapa são utilizadas 4 ferramentas de Unitex, nessa ordem: `Normalize`, `Tokenize`, `Dico` e `SortTxt`. Além disso, antes da realização da primeira etapa, ou seja, antes da execução de `Normalize`, o arquivo a ser processado é convertido da codificação de caracteres ISO-8859-1 (codificação do arquivo de entrada) para Unicode (codificação usada por Unitex) e, após a execução de `SortTxt`, o processo reverso é executado convertendo o arquivo de Unicode para ISO-8859-1.

`Normalize` é responsável por realizar uma normalização dos separadores de texto: espaço, tabulação e caractere de mudança de linha. Cada seqüência de separadores que contém pelo menos um caractere de mudança de linha é substituída

por um único caractere de mudança de linha. Todas as outras seqüências de separadores são substituídas por um único espaço. Esse programa também substitui seqüências de caracteres "{" e "}" – que têm significado especial para Unitex – por "([" e "])". A saída é um arquivo com a extensão ".snt" cujo conteúdo é uma versão modificada do texto.

Em seguida, `Tokenize` separa o texto contido no .snt gerado por `Normalize` em *tokens* de acordo com o alfabeto definido no arquivo Alphabet.txt. O arquivo com o alfabeto é distribuído juntamente com Unitex e deve estar presente no subdiretório "unitex" contido no diretório de dados do idioma passado como parâmetro para o programa (`-d`). Entre os arquivos produzidos como saída de `Tokenize` estão tokens.txt (arquivo com a lista de *tokens* identificados) e text.cod (arquivo binário com a seqüência de códigos que representam cada *token*) essenciais para o restante do processamento. Todos os arquivos gerados por `Tokenize` são armazenados em um diretório auxiliar criado apenas para essa etapa e removido após seu término.

A próxima ferramenta, `Dico`, é a mais importante para o processamento em questão pois é a responsável por atribuir a cada palavra suas possíveis análises. Para tanto, `Dico` consulta os dicionários de palavras simples e compostas de Unitex. Esses dicionários são distribuídos juntamente com Unitex e também devem estar presentes no subdiretório "unitex". Os dicionários devem ser arquivos .bin (obtidos com o programa Compress de Unitex) ou um grafo de dicionário no formato .fst2. Além disso, é possível especificar qual dicionário deve ser aplicado primeiro (detalhes podem ser obtidos em (PAUMIER, 2006)). No caso da ferramenta apresentada neste relatório, apenas o dicionário de palavras simples é aplicado e, portanto, para sua execução, apenas os arquivos delaf.bin e delaf.inf devem estar contidos no subdiretório "unitex".

Como saída, `Dico` produz quatro arquivos (os quais também são salvos no diretório criado temporariamente para esta etapa): (1) dicionário das palavras simples no texto (dlf), (2) dicionário das palavras compostas no texto (dlc), (3) lista de palavras desconhecidas no texto (err) e (4) arquivo contendo o número de

palavras simples, compostas e desconhecidas no texto (stat_dic.n). Desses quatro arquivos o único que é processado por `anali` é o primeiro, dlf, o qual é ordenado pelo programa `SortTxt` de acordo com a ordem lexicográfica dos caracteres Unicode.

Ao final da primeira etapa e após a conversão do arquivo dlf de Unicode para ISO-8859-1, todas as análises são lidas e armazenadas numa estrutura de dados do tipo vetor associativo (*hash*) utilizada no restante do processamento.

Na segunda etapa, de maneira semelhante ao modo `e`, o arquivo de entrada é pré-processado para inserir espaços em lugares pré-estabelecidos. Em seguida, o vetor associativo com as análises retornadas por Unitex é consultado e a cada palavra são atribuídas suas possíveis análises delimitadas pelos caracteres "{" e "}".

Por padrão, o arquivo de saída gerado no modo `a` possui o mesmo nome do arquivo de entrada, porém com a extensão ".a". Além disso, de maneira semelhante ao que acontece no modo `e`, se o arquivo de entrada estiver etiquetado com fronteiras de sentenças essas etiquetas também estarão presentes no arquivo de saída.

Assim, a Tabela 4 apresenta um trecho de um arquivo em inglês analisado por meio do comando:

```
> perl anali.pl -m a -a ingles_et.txt -d Ingles
```

O texto armazenado em ingles_et.txt é o mesmo do texto ingles.txt usado como exemplo de etiquetação apresentado na Tabela 2, porém com etiquetas delimitadoras de sentença (<s snum=x> e </s>).

O texto em português (portugues_et.txt) , também etiquetado com delimitadores de sentenças (<s snum=x> e </s>) correspondente ao conteúdo do texto etiquetado anteriormente (veja Tabela 3) é analisado por meio da execução do comando:

Tabela 4: Três primeiras sentenças de um texto em inglês analisado por `anali`

<s snum=1>Re{.N:s:p}{.PFX} - inventing{invent.V:G} Benjamin{.N+Hum:s}{.N+Conc:s} Franklin{.N+Hum:s}</s>

<s snum=2>A{.DET+Dind:s}{.N:s} team{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} of{.PREP} Brazilian{.A}{.N+Hum:s} researchers{researcher.N+Hum:p} produced{.A}{produce.V:K:I1s:I2s:I3s:I1p:I2p:I3p} artificial{.A} lightning{.A}{.N:s}{.V:W:G} for{.CONJ}{.PREP} the{.DET+Ddéf:s:p} first{.A}{.ADV}{.N:s} time{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} in{.A}{.PREP}{.N:s}{.PART} the{.DET+Ddéf:s:p} Southern{.A}{.N:s} Hemisphere{.N:s} .</s>
<s snum=3>In{.A}{.PREP}{.N:s}{.PART} the{.DET+Ddéf:s:p} final{.A}{.N:s} days{.ADV}{day.N+Ntime:p} of{.PREP} November{.N+Ntime:s} , the{.DET+Ddéf:s:p} people{.N+Hum:s}{.N+HumColl:p}{.V:W:P1s:P2s:P1p:P2p:P3p} at{.PREP} the{.DET+Ddéf:s:p} International{.A}{.N+Hum:s} Center{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} of{.PREP} Research{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} and{.V+i:W:P1s:P2s:P1p:P2p:P3p}{.CONJ} Tests{test.N:p}{test.V:P3s} on{.A}{.PREP}{.PART} Lightning{.A}{.N:s}{.V:W:G} , of{.PREP} the{.DET+Ddéf:s:p} National{.A}{.N+Hum:s} Institute{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} of{.PREP} Spatial{.A} Research{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} (Inpe) , repeated{.A}{repeat.V:K:I1s:I2s:I3s:I1p:I2p:I3p} the{.DET+Ddéf:s:p} experiment{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} done{.A}{do.V:K} some{.PRO:s:p}{.DET+Dadj:s:p} 248 years{year.N+Ntime:p} ago{.ADV} when{.CONJ}{.ADV}{.N:s} Benjamin{.N+Hum:s}{.N+Conc:s} Franklin{.N+Hum:s} flew{.N:s}{fly.V:I1s:I2s:I3s:I1p:I2p:I3p} a{.DET+Dind:s}{.N:s} kite{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} with{.PREP} a{.DET+Dind:s}{.N:s} spanner{.N+Conc:s} on{.A}{.PREP}{.PART} the{.DET+Ddéf:s:p} end{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} of{.PREP} it{.PRO:3ns} to{.PREP}{.PART} prove{.V:W:P1s:P2s:P1p:P2p:P3p} that{.CONJ}{.ADV}{.DET+Ddem:s}{.PRO+Pdem:s} the{.DET+Ddéf:s:p} atmosphere{.N:s} was{be.V:I1s:I3s} full{.A}{.ADV}{.N:s}{.V:W:P1s:P2s:P1p:P2p:P3p} of{.PREP} electricity{.N:s} .</s>
...

```
> perl anali.pl -m a -a portugues_et.txt -d Portugues
```

Parte do texto analisado produzido como saída é apresentada na Tabela 5.

2.3 Intersecção entre etiquetação e análise

No último modo de processamento (`ae`), `anali` utiliza a saída de MXPOST (e opcionalmente algumas regras de desambiguação) para determinar qual das

Tabela 5: Três primeiras sentenças de um texto em português analisado por `anali`

<s snum=1>Reinventando{reinventar.V:G} Benjamin{.N+Pr:ms} Franklin{.N+Pr:ms:fs}</s>

<s snum=2>Uma{um.A:fs}{um.PRO+Ind:fs}{um.DET+Num:Cfs}{um.DET+Art+Ind:fs} equipe{.N:fs}{equipar.V:S1s:S3s:Y3s} de{.PREP} pesquisadores{pesquisador.A:mp}{pesquisador.N:mp} brasileiros{brasileiro.A:mp}{brasileiro.N:mp} produziu{produzir.V:J3s} raios{raio.N:mp} artificialmente{.ADV} pela{pelo.PREPXDET+Art+Def:fs}{pelo.PREPXPRO+Dem:fs} primeira{primeiro.A:fs}{primeiro.DET+Num:Ofs} {primeiro.N:fs} vez{.N:fs} no{ele.PRO+Pes:O3ms:A3ms:D3ms}{.PREPXDET+Art+Def:ms} {.PREPXPRO+Dem:ms} Hemisfério{.N:ms} Sul{.A:ms:fs}{.N:ms} .</s>
<s snum=3>Nos{no.PREPXDET+Art+Def:mp}{ele.PRO+Pes:O3mp:A3mp:D3mp|eu.PRO+Pes:O1mp:A1mp:D1mp:R1mp:O1fp:A1fp:D1fp:R1fp}{no.PREPXPRO+Dem:mp} últimos{último.A:mp}{último.N:mp} dias{dia.N:mp}{.N+Pr:ms:mp:fs:fp} de{.PREP} novembro{.N:ms} , o{.PRO+Dem:ms}{.N:ms}{ele.PRO+Pes:A3ms} {.DET+Art+Def:ms} pessoal{.A:ms:fs}{.N:ms:fs} do{.PREPXDET+Art+Def:ms} {.PREPXPRO+Dem:ms} Centro{.N:ms}{centrar.V:P1s} Internacional{.A:ms:fs}{.N:ms:fs} de{.PREP} Pesquisa{.N:fs}{pesquisar.V:P3s:Y2s} e{.CONJ}{.N:ms} Testes {teste.N:mp}{testar.V:S2s} de{.PREP} Raios{raio.N:mp} , do{.PREPXDET+Art+Def:ms}{.PREPXPRO+Dem:ms} Instituto{.N:ms} Nacional{.A:ms:fs}{.N:ms:fs} de {.PREP} Pesquisas{pesquisa.N:fp}{pesquisar.V:P2s} Espaciais{espacial.A:mp:fp {espaciar.V:P2p} (Inpe{.SIGL}) , repetiu{repetir.V:J3s} a{.PREP}{o.PRO+Dem:fs}{.N:ms}{ele.PRO+Pes:A3fs}{o.DET+Art+Def:fs}{.ABREV:ms} experiência{.N:fs} feita{feito.A:fs}{feito.N:fs}{fazer.V:K|feitar.V:P3s:Y2s} há{haver.V:P3s:Y2s} 248 anos{ano.N:mp} , quando{.CONJ}{.ADV}{.PRO+Rel:ms:mp:fs:fp} Benjamin{.N+Pr:ms} Franklin{.N+Pr:ms:fs} empinou{empinar.V:J3s} uma{um.A:fs}{um.PRO+Ind:fs}{um.DET+Num:Cfs}{um.DET+Art+Ind:fs} pipa{.N:fs} com{.PREP} {.ABREV:ms} uma{um.A:fs} {um.PRO+Ind:fs}{um.DET+Num:Cfs}{um.DET+Art+Ind:fs} chave{.N:fs} na{no.PREPXDET+Art+Def:fs}{ele.PRO+Pes:O3fs:A3fs:D3fs} {no.PREPXPRO+Dem:fs} extremidade{.N:fs} para{.PREP}{.PFX}{parir.V:Y3s} provar{.V:W1s:W3s:U1s:U3s} que{.CONJ}{.ADV}{.PRO+Ind:ms:mp:fs:fp}{.PRO+Int:ms:mp:fs:fp}{.PRO+Rel:ms:mp:fs:fp} a{.PREP}{o.PRO+Dem:fs}{.N:ms}{ele.PRO+Pes:A3fs}{o.DET+Art+Def:fs}{.ABREV:ms} atmosfera{.N:fs} estava{estar.V:I1s:I3s} carregada{carregado.A:fs}{carregado.N:fs}{carregar.V:K} de{.PREP} eletricidade{.N:fs} .</s>
...

análises retornadas por Unitex é a melhor para cada palavra. Para tanto, `anali` primeiro realiza as mesmas tarefas dos modos de etiquetação (veja seção 2.1) e análise (veja seção 2.2) apresentadas anteriormente. Em seguida, outras três ta-

refas são realizadas com o intuito de retornar, para cada *token*, o maior conjunto de informações que seja comum às saídas de MXPOST e Unitex: (1) mapeamento das etiquetas, (2) intersecção das análises e, opcionalmente, (3) desambiguação das análises com base em regras definidas manualmente.

O mapeamento das etiquetas de MXPOST para as etiquetas de Unitex é realizado para possibilitar a intersecção de suas saídas, já essas ferramentas não utilizam o mesmo conjunto de etiquetas (veja Anexos A e B). Para tanto, o modo `ae` exige como parâmetro de entrada um arquivo contendo o mapeamento das etiquetas de MXPOST para as etiquetas de Unitex no qual cada linha deve obedecer o seguinte formato:

$\texttt{ETIM}_i\texttt{=ETIU}_j\texttt{(|ETIU}_k\texttt{)*}$

em que ETIM_i é uma etiqueta de MXPOST e ETIU_n é sua correspondente no Unitex. Caso uma etiqueta de MXPOST corresponda a mais de uma etiqueta no Unitex, as etiquetas de Unitex devem ser enumeradas separadas pelo caractere "|". Cada mapeamento de uma etiqueta de MXPOST para uma ou mais etiquetas de Unitex deve ser especificado em uma linha separada do arquivo de mapeamento. Comentários são inseridos em linhas iniciadas com o caractere "%".

Assim, por exemplo, a linha "IN=PREP|CONJS" (presente no arquivo de mapeamento definido para o idioma inglês, veja Apêndice A) estabelece que a etiqueta "IN" de MXPOST pode ser mapeada, no Unitex, como uma preposição (PREP) ou uma conjunção de subordinação (CONJS). O mapeamento pode, ainda, envolver etiquetas que representam, além da categoria gramatical, também seus atributos; como acontece no mapeamento da etiqueta de MXPOST para adjetivos comparativos do inglês (JJR) e a combinação de etiqueta e atributo do Unitex (A:C): "JJR=A:C".

Os arquivos com os mapeamentos das etiquetas de MXPOST para as etiquetas de Unitex definidos para os idiomas inglês e português estão presentes no Apêndice A. Esses arquivos foram criados com base na análise dos conjuntos de

etiquetas de MXPOST (veja Anexo A) e de Unitex (veja Anexo B) e em alguns estudos realizados com corpus.

Após o mapeamento das etiquetas de MXPOST para Unitex tanto a de saída do modo de etiquetação quanto a saída do modo de análise são representadas com etiquetas de Unitex e podem, assim, serem comparadas com o intuito de determinar o maior conjunto de informações morfossintáticas em comum entre as saídas de MXPOST e de Unitex.

A próxima tarefa, a intersecção, na verdade pode ser entendida como um processo de desambiguação da saída de Unitex ou de enriquecimento da saída de MXPOST. A desambiguação da saída de Unitex ocorre uma vez que MXPOST ajuda a selecionar a melhor análise entre todas as análises retornadas por Unitex. Enquanto o enriquecimento da saída de MXPOST é o resultado da união das informações morfológicas e semânticas disponíveis nas análises de Unitex com a categoria gramatical (e, às vezes, alguns poucos atributos morfossintáticos) retornada por MXPOST. Nessa intersecção, as análises retornadas por MXPOST e por Unitex são comparadas e apenas aquelas que são comuns (mesma categoria gramatical e, possivelmente, alguns traços morfossintáticos) são mantidas.

Por fim e opcionalmente, a última tarefa realizada em busca da melhor intersecção entre MXPOST e Unitex tem como base algumas regras de desambiguação especificadas em um arquivo passado como parâmetro de entrada (`-des`). Essas regras de desambiguação devem obedecer a um dos três formatos apresentados a seguir:

1. CAT
   n [CAT]
2. CAT
   n {comandos}
3. CAT
   n “*token*”

Todos os três formatos possuem dois elementos comuns: CAT e n. CAT é uma categoria gramatical de acordo com o conjunto de etiquetas de Unitex a qual a regra se aplica, ou seja, aquela que será retornada como a melhor categoria gramatical se a condição definida na regra for satisfeita; e n determina qual a posição da palavra que deve ser analisada, em relação à posição da palavra sendo desambiguada. Cada um desses três formatos é apresentado em detalhes a seguir.

O primeiro formato pode ser usado para definir uma regra de desambiguação que analisa apenas as categorias gramaticais das palavras vizinhas à palavra que se deseja desambiguar. Por exemplo, a regra utilizada para definir se uma palavra em inglês (veja regras de desambiguação no Apêndice B) é uma partícula (PART) pode ser descrita em termos da categoria gramatical de sua antecessora: se a categoria gramatical de sua antecessora ($n = -1$) for verbo (V) então a palavra é uma partícula. Tal regra é definida, usando o primeiro formato, como:

PART

-1 [V]

O segundo formato utiliza comandos da linguagem Perl para verificar se CAT é a melhor categoria gramatical. Esses comandos devem, necessariamente, retornar um de dois valores: verdadeiro (1) ou falso (0). Os comandos são aplicados ao elemento deslocado n posições da palavra a ser desambiguada e se o resultado for verdadeiro (1) a categoria CAT é considerada a melhor. Esse formato de regra é usado, por exemplo, para definir a regra de desambiguação de um nome próprio em inglês (N+PR) definida como: se a palavra ($n = 0$) começa com letra maiúscula, então ela é um nome próprio.[3] Essa regra é definida, seguindo o segundo formato, como:

N+PR

0 {ucfirst($_) eq $_}

[3]Uma regra de desambiguação semelhante também foi definida para nomes próprios em português: N+Pr.

Por fim, o último formato previsto para uma regra de desambiguação permite uma comparação lexical, ou seja, a categoria CAT será eleita a melhor se a palavra deslocada n posições da palavra a ser desambiguada for igual a *token*. Por exemplo, esse formato é usado para a regra aplicada aos prefixos em inglês ou português (PFX): se a sucessora ($n = 1$) da palavra a ser desambiguada é o *token* "-", então ela é um prefixo. Tal regra é definida, de acordo com o terceiro formato, como:

PFX

1 "-"

As três regras de desambiguação apresentadas anteriormente podem ser utilizadas por `anali` no processamento de textos em inglês e estão contidas no arquivo apresentado no Apêndice B, seção B.1. Um arquivo semelhante foi gerado para o processamento de textos em português, o qual pode ser consultado em Apêndice B, seção B.2.

Com o intuito de esclarecer como o modo `ae` desambigua/enriquece as saídas de Unitex/MXPOST, alguns exemplos do resultado desse modo são apresentados na Tabela 6 para palavras em inglês.

Os primeiros dois exemplos são de *tokens* que foram etiquetados por MXPOST porém não foram analisados por Unitex pois não estão presentes no seu dicionário de palavras simples já que se trata de uma sigla do português (e_1) e um número (e_2). Nesse caso, a saída do modo `ae` é considerada a mesma de MXPOST.

O terceiro exemplo, e_3, demonstra claramente a desambiguação das análises de Unitex (enriquecimento da saída de MXPOST) com base na saída de MXPOST (Unitex) uma vez que para MXPOST a palavra *a* é apenas um determinante (DET) e, para Unitex, ela pode ser um determinante indefinido singular (DET+Dind:s) ou um substantivo singular (N:s). A saída, nesse caso, é a melhor combinação entre a saída das duas ferramentas o que resulta na análise de *a* como um determinante indefinido singular (DET+Dind:s).

Os dois exemplos seguintes (e_4 e e_5) representam casos nos quais o etiquetador

Tabela 6: Exemplos da saída do modo `ae`

#	**Unitex** (`a`)	**MXPOST** (`e`)	**Intersecção** (`ae`)
e_1	Inpe	Inpe_N+PR:s	Inpe/.N+PR:s
e_2	248	248_DET+Dnum	248/.DET+Dnum
e_3	a{.DET+Dind:s}{.N:s}	a_DET	a/.DET+Dind:s
e_4	artificial{.A}	artificial_A	artificial/.A
e_5	inventing{invent.V:G}	inventing_V:G	inventing/invent.V:G
e_6	Hemisphere{.N:s}	Hemisphere_N+PR:s	Hemisphere/.N:s
e_7	days{.ADV}{day.N+Ntime:p}	days_N:p	days/day.N+Ntime:p
e_8	November{.N+Ntime:s}	November_N+PR:s	November/.N:s
e_9	copper{.N+Hum:s}{.N+Conc:s} {.V:W:P1s:P2s:P1p:P2p:P3p}	copper_N:s	copper/.N:s
e_{10}	people{.N+Hum:s}{.N+HumColl :p}{.V:W:P1s:P2s:P1p:P2p:P3p}	people_N:p	people/.N+HumColl:p
e_{11}	prove{.V:W:P1s:P2s:P1p:P2p:P3p}	prove_V:W	prove/.V:W
e_{12}	is{i.N:p}{be.V:P3s}	is_V:P3s:S3s	is/be.V:P3s
e_{13}	are{.N+Unit:s} {be.V:P2s:P1p:P2p:P3p}	are_V:P1s:P2s:S1s :S2s	are/be.V:P2s
e_{14}	the{.DET+Ddéf:s:p}	the_DET	the/.DET+Ddéf
e_{15}	Spatial{.A}	Spatial_N+PR:s	Spatial/.N+PR:s
e_{16}	up{.A}{.PREP}{.N:s}{.PART} {.V:W:P1s:P2s:P1p:P2p:P3p}	up_ADV|ADVA	up/.PART
e_{17}	to{.PREP}{.PART}	to_PREP|PART	to/.PREP
e_{18}	that{.CONJ}{.ADV} {.DET+Ddem:s}{.PRO+Pdem:s}	that_PREP|CONJS	that/.PREP

e o analisador estão de acordo com qual categoria devem atribuir à palavra e só há uma categoria possível. No exemplo (e_5) além de concordarem em relação à categoria gramatical da palavra (verbo, V), também concordam em termos dos atributos (o verbo está no gerúndio, G). Nem sempre a concordância entre as duas ferramentas resulta em uma única análise como é o caso do exemplo e_{17} apresentado a seguir.

Além disso, para tornar o processo ainda mais complexo, uma categoria gramatical de Unitex pode vir acompanhada de traços semânticos/morfológicos como ocorre, por exemplo, nos exemplos e_6 a e_9. Nesses casos, três regras são aplicadas para se determinar se os traços devem ser incluídos na categoria gramatical

atribuída à palavra em questão: (1) se a versão etiquetada possui traços mas a analisada não, os traços não são incluídos na saída (veja e_6); (2) por outro lado, se a versão etiquetada não possui traços mas a analisada sim e estes são únicos, então os traços são incluídos na saída (veja e_7) por outro lado, se houver mais de uma possibilidade de traços, nenhuma é incluída na saída (veja e_9); por fim, (3) se tanto a versão etiquetada quanto a analisada possuem traços mas eles são diferentes, então nenhum traço é inserido na saída (veja e_8).

Os próximos quatro exemplos demonstram como os atributos da versão etiquetada (e_{10} e e_{11}), analisada (e_{12}) ou ambos (e_{13}) podem ser usados para determinar o melhor conjunto de informações mesmo quando não aparecem sozinhos, como é o caso do exemplo e_{13} no qual apenas um conjunto de atributos é obtido na intersecção das várias opções tanto na versão analisada quanto na etiquetada.

Porém, quando não é possível selecionar um único conjunto de atributos, as várias opções não são incluídas na saída para se evitar a inserção de ambigüidades como ocorre com o atributo de número no exemplo e_{14}.

Os três exemplos seguintes são obtidos como resultado da aplicação das três regras de desambiguação apresentadas anteriormente. No exemplo e_{15} não há concordância a respeito da categoria gramatical da palavra *Spatial* nas versões analisada e etiquetada e, portanto, todas as possíveis análises são avaliadas pelas regras de desambiguação. Nesse caso, a regra definida anteriormente para a categoria nome próprio (N+PR) tem sucesso uma vez que a palavra em questão (*Spatial*) é iniciada com letra maiúscula. No exemplo e_{16} também não há intersecção entre as categorias das versões analisada e etiquetada e, nesse caso, a primeira regra de desambiguação apresentada anteriormente especifica que a palavra deve ser classificada como partícula já que sua antecessora foi analisada como "went/go.V", ou seja, um verbo.

Outro fato importante a respeito do processo de desambiguação é que mesmo que uma regra não tenha sucesso ela pode, sim, levar à seleção da melhor categoria. Isso ocorre porque toda vez que uma regra é aplicada para uma determinada

categoria e não tem sucesso, essa categoria é removida do conjunto de possíveis categorias reduzindo-se, assim, a ambigüidade. O exemplo e_{17} apresenta uma palavra que foi desambiguada com a falha de uma regra; trata-se da análise da palavra *to*, nesse caso, precedida por "it/.PRO:3ns". Como a antecessora de *to* não é um verbo mas sim um pronome, a palavra não pode ser categorizada como partícula; a categoria PART é removida do conjunto de possibilidades restando, apenas, a categoria PREP que é, então, atribuída à palavra em questão.

Se ao final de todo o processo de desambiguação ainda existir mais de uma análise possível para uma dada palavra, `anali` considerará todas as análises ou apenas a primeira delas dependendo da escolha do usuário. Se o usuário especificar, como parâmetro de entrada do programa, que deseja que todas as análises possíveis sejam mantidas (opção `-t`) elas serão impressas na saída, caso contrário apenas a primeira opção será impressa. É importante citar, aqui, que `anali` prioriza as análises retornadas pelo etiquetador e considera a ordem na qual as etiquetas são definidas no arquivo de mapeamento. O exemplo e_{18} apresenta um caso em que a primeira opção retornada pelo etiquetado (PREP) foi selecionada considerando-se que os processos de intersecção e desambiguação não tiveram êxito e a opção `-t` não foi especificada.

Assim, a Tabela 7 apresenta um trecho de um arquivo em inglês etiquetado e analisado por meio do comando:

```
> perl anali.pl -m ae -a ingles_et.txt -d Ingles -map
MXPOST_Unitex_en.txt -des regras_desamb_en.txt -t
```

O texto de entrada, ingles_et.txt, é o mesmo apresentado nas Tabelas 2 (como saída do modo `e`) e 4 (como saída do modo `a`) também com etiquetas delimitadoras de sentença (<s snum=x> e </s>). Além disso, os arquivos de mapeamento e de regras de desambiguação especificados como parâmetro de entrada nesse comando podem ser consultados nos Apêndices A e B, respectivamente. Por fim, a opção `-t` no comando especifica que todas as análises possíveis de uma palavra deverão ser mantidas quando os processos de intersecção e desambiguação não forem capazes

de apontar uma única análise.

Tabela 7: Três primeiras sentenças de um texto em inglês analisado e etiquetado por `anali`

<s snum=1>Re/.N:s - inventing/invent.V:G Benjamin/.N:s Franklin/.N:s</s>

<s snum=2>A/.DET+Dind:s team/.N:s of/.PREP Brazilian/.A researchers/researcher.N+Hum:p produced/produce.V artificial/.A lightning/.N:s for/.PREP the/.DET+Ddéf first/.A time/.N:s in/.PREP the/.DET+Ddéf Southern/.N:s Hemisphere/.N:s .</s>
<s snum=3>In/.PREP the/.DET+Ddéf final/.A days/day.N+Ntime:p of/.PREP November/.N:s , the/.DET+Ddéf people/.N+HumColl:p at/.PREP the/.DET+Ddéf International/.N:s Center/.N:s of/.PREP Research/.N:s and/.CONJ Tests/test.V:P3s on/.PREP Lightning/.N:s , of/.PREP the/.DET+Ddéf National/.N:s Institute/.N:s of/.PREP Spatial/.N+PR:s Research/.N:s (Inpe/.N+PR:s) , repeated/repeat.V the/.DET+Ddéf experiment/.N:s done/do.V:K some/.DET+Dadj 248/.DET+Dnum years/year.N+Ntime:p ago/.ADV when/.ADV Benjamin/.N:s Franklin/.N:s flew/fly.V a/.DET+Dind:s kite/.N:s with/.PREP a/.DET+Dind:s spanner/.N+Conc:s on/.PREP the/.DET+Ddéf end/.N:s of/.PREP it/.PRO:3ns to/.PREP prove/.V:W that/{.CONJ}{.PREP}{.CONJS}{.ADV}{.DET+Ddem:s}{.PRO+Pdem:s}
the/.DET+Ddéf atmosphere/.N:s was/be.V full/.A of/.PREP electricity/.N:s .</s>
...

De maneira semelhante, a Tabela 8 apresenta um trecho de um arquivo em português etiquetado e analisado por meio do comando

```
> perl anali.pl -m ae -a portugues_et.txt -d Portugues -map
MXPOST_Unitex_pt.txt -des regras_desamb_pt.txt
```

O texto de entrada, portugues_et.txt, é o mesmo apresentado nas Tabelas 3 (como saída do modo `e`) e 5 (como saída do modo `a`) também com etiquetas delimitadoras de sentença (<s snum=x> e </s>). Além disso, os arquivos de mapeamento e de regras de desambiguação especificados como parâmetro de entrada nesse comando podem ser consultados nos Apêndices A e B, respectivamente. Diferentemente do exemplo em inglês, neste caso a opção `-t` não é passada como parâmetro e, portanto, em caso de mais de uma análise possível, apenas a primeira será considerada.

Tabela 8: Três primeiras sentenças de um texto em português analisado e etiquetado por `anali`

<s snum=1>Reinventando/.N+Pr Benjamin/.N+Pr:ms Franklin/.N+Pr</s>

<s snum=2>Uma/um.DET+Art+Ind:fs equipe/.N:fs de/.PREP pesquisadores/pesquisador.N:mp brasileiros/brasileiro.A:mp produziu/produzir.V:J3s raios/raio.N:mp artificialmente/.ADV pela/pelo.PREPXDET+Art+Def:fs primeira/primeiro.A:fs vez/.PREP no/.PREPXDET+Art+Def:ms Hemisfério/.N:ms Sul/.N:ms .</s>
<s snum=3>Nos/ele.PRO últimos/último.A:mp dias/dia.N:mp de/.PREP novembro/.N:ms , o/.DET+Art+Def:ms pessoal/.N do/.PREPXDET+Art+Def:ms Centro/.N:ms Internacional/.N de/.PREP Pesquisa/.N:fs e/.CONJ Testes/teste.N:mp de/.PREP Raios/raio.N:mp , do/.PREPXDET+Art+Def:ms Instituto/.N:ms Nacional/.N de/.PREP Pesquisas/pesquisa.N:fp Espaciais/.N+Pr (Inpe/.N+Pr) , repetiu/repetir.V:J3s a/o.DET+Art+Def:fs experiência/.N:fs feita/fazer.V:K há/.PREP 248/.DET+Num anos/ano.N:mp , quando/.ADV Benjamin/.N+Pr:ms Franklin/.N+Pr empinou/empinar.V:J3s uma/um.DET+Art+Ind:fs pipa/.N:fs com/.PREP uma/um.DET+Art+Ind:fs chave/.N:fs na/no.PREPXDET+Art+Def:fs extremidade/.N:fs para/.PREP provar/.V que/.PRO+Ind a/o.DET+Art+Def:fs atmosfera/.N:fs estava/estar.V carregada/carregar.V:K de/.PREP eletricidade/.N:fs .</s>
...

3 Considerações finais

Este relatório apresentou a ferramenta de análise morfossintática `anali` desenvolvida no NILC como o resultado da união de duas outras ferramentas de Processamento de Língua Natural: o etiquetador MXPOST (RATNAPARKHI, 1996) e a ferramenta de processamento de corpus Unitex (PAUMIER, 2006).

A nova ferramenta representa um ganho em relação ao que é produzido pelas outras duas em dois sentidos: por um lado, enriquece a saída de MXPOST inserindo mais informações à análise retornada por este etiquetador; e, por outro, desambigua a saída de Unitex já que tenta definir qual das várias análises existentes em seus dicionários eletrônicos é a melhor.

Além das ferramentas citadas, `anali` pode se basear em regras de desambiguação, criadas manualmente e definidas em um arquivo de entrada, para auxiliar no processo de seleção da melhor análise.

Embora este relatório apresente apenas arquivos de mapeamento e regras para os idiomas inglês e português, `anali` foi projetada para ser usada com qualquer idioma para o qual estejam disponíveis modelos treinados de MXPOST e dicionários eletrônicos de Unitex bastando, apenas, criar/adaptar os arquivos de mapeamento e de regras de desambiguação para esses novos idiomas.

Referências

AIRES, R. V. X. *Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP, 2000.

ALUÍSIO, S. et al. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: *Lecture Notes on Artificial Intelligence. Proceedings of PROPOR 2003*. Springer-Verlag, 2003. v. 1. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/publicacoes.htm>.

BRILL, E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, v. 21, n. 4, p. 543–565, 1995. Disponível em: <http://www.cs.jhu.edu/~brill/>.

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, v. 19, n. 2, p. 313–330, 1993.

MUNIZ, M. C. M. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP, 2004. Disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/publicacoes.html>.

PAUMIER, S. *Unitex 1.2 User Manual*. [S.l.], June 2006. Disponível em: <http://www-igm.univ-mlv.fr/~unitex/>.

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: *Proceedings of the First Empirical Methods in NLP Conference*. [S.l.: s.n.], 1996.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK: [s.n.], 1994. p. 44–49.

APÊNDICE A – Mapeamento das etiquetas de MXPOST para as de Unitex

A seguir são apresentados os arquivos com os mapeamentos entre as etiquetas de MXPOST e Unitex para os idiomas inglês (seção A.1) e português (seção A.2).

No mapeamento das etiquetas para o idioma inglês, foram criadas duas novas etiquetas não existentes em Unitex com o intuito de evitar a perda de informação proveniente de MXPOST, são elas: `CUR` para marcação de símbolos de moeda ($, R$) e `POS` para marcação de terminação de possessivo ('s). De maneira semelhante, no mapeamento para o idioma português foram criadas três novas etiquetas: `KS` (conjunção subordinativa), `VAUX` (verbo auxiliar) e `CUR` (símbolos de moeda).

A.1 Mapeamento para o inglês

Conteúdo do arquivo MXPOST_Unitex_en.txt

% TagSet MXPOST: Penn Tagset (`http://www.mozart-oz.org/mogul/doc/lager/brill-tagger/penn.html`)
% Unitex: categorias definidas para o ingles (Paumier, 2006)

% Etiquetas existentes em MXPOST mas nao em Unitex: POS, CUR

```
% Etiquetas de MXPOST nao mapeadas: LS, SYM

%MXPOST=Unitex

% conjuncao coordenativa: and
CC=CONJ

% | IN: preposicao/conjuncao subordinativa: in, of, like, because
IN=PREP|CONJS

% numeral: 1, third
CD=DET+Dnum

% determinante: the
DT=DET

% | there: there is (Unitex: there,.ADV or there,.INTJ or there,.PRO+Nomin:3ns:3np)
EX=ADV|INTJ|PRO

% | palavra estrangeira: d'hoevre
FW=X|XI

% adjetivo: green
JJ=A

% adjetivo comparativo: greener
JJR=A:C

% adjetivo superlativo: greenest
JJS=A:S
```

% verbo modal: could, will
MD=V+aux

% substantivo singular ou de massa: table
NN=N:s

% substantivo plural: tables
NNS=N:p

% nome proprio singular: John
NNP=N+PR:s

% nome proprio plural: Vikings
NNPS=N+PR:p

% pre-determinante: both (the boys)
PDT=PRED

% MXPOST terminacao de possessivo: friend's. Nao ha etiqueta para possessivo no Unitex, mas para nao perder essa informacao, cria-se uma
POS=POS

% | pronome pessoal: I, he, it (PRON = whatever)
PRP=PRO|PRON

% | pronome possessivo: my, his
PRP$=PRO:Poss1s|PRO:Poss1p|PRO:Poss2sp|PRO:Poss3fs|PRO:Poss3ms|PRO:Poss3ns|PRO:Poss3p

% | pronome-wh possessivo: whose

WP$=PRO:Poss1s|PRO:Poss1p|PRO:Poss2sp|PRO:Poss3fs|PRO:Poss3ms|PRO:Poss3ns|PRO:Poss3p

% | adverbio: however, usually, naturally, here, good (ADVA = somewhat)
RB=ADV|ADVA

% adverbio comparativo: better
RBR=ADV:C

% adverbio superlativo: best
RBS=ADV:S

% particula: (give) up
RP=PART

% | to (go), to (him)
TO=PREP|PART

% interjeicao: uhhuhhuhh
UH=INTJ

% verbo no infinitivo: take
VB=V:W

% verbo no passado: took
VBD=V:J:I:T

% verb no gerundio/participio presente: taking
VBG=V:G

% verbo no participio, passsado: taken

VBN=V:K

% verbo no presente cuja pessoa nao eh a terceira do singular: take
VBP=V:P1s:P2s:S1s:S2s

% verbo no presente cuja pessoa eh a terceira do singular: takes
VBZ=V:P3s:S3s

% determinante-wh: which
WDT=DET+DetQ

% pronome-wh: who, what
WP=PRO+RelQ

% adverbio-wh: where, when
WRB=ADV

% simbolos de moeda. Nao ha etiqueta para simbolos de moeda no Unitex, mas para nao perder essa informacao, cria-se uma
$=CUR

A.2 Mapeamento para o português

Conteúdo do arquivo MXPOST_Unitex_pt.txt

% TagSet MXPOST: artigo do PROPOR (Aluisio et al., 2003) e diretrizes para a etiquetacao manual do corpus Mac-Morpho em `http://www.nilc.icmc.usp.br/lacioweb/manuais.htm`
% Unitex: etiquetas definidas para o portugues (Muniz, 2004)

% Etiquetas existentes em MXPOST mas nao em Unitex: KS, VAUX, CUR
% Etiqueta de MXPOST nao mapeada: PDEN

%MXPOST=Unitex

% adjetivo: bonito, bonitas, simples
ADJ=A

% adverbio: abaixo, mesmo (no Unitex, nao ha atributos para ADV)
ADV=ADV
ADV-KS-REL=ADV
ADV-KS=ADV

% artigo: o, umas
ART=DET+Art

% conjuncao coordenativa: mais, mas, mal
KC=CONJ

% conjuncao subordinativa: que, embora, se. Nao ha etiqueta para conjuncao subordinativa no Unitex, mas para nao perder essa informacao, cria-se uma
KS=KS

% interjeicao: ah, ih, oi
IN=INTERJ

% substantivo: menino, meninos
N=N

% nome proprio: Ana, Silva

NPROP=N+Pr

% numeral: segundo, duplo, oito
NUM=DET+Num

% | participio passado ou adjetivo: realizado, privadas
PCP=V:K|A

% preposicao: ante, de
PREP=PREP

% pronome pessoal: eu
PROPESS=PRO+Pes

% pronome subordinado relativo: quando
PRO-KS-REL=PRO+Rel

% pronome subordinado nao-relativo: que, quantos
PRO-KS=PRO

% pronome nao-subordinado como nucleo de um sintagma nominal: nenhum, poucos
PROSUB=PRO

% pronome não-subordinado como um modificador: poucas, ambos
PROADJ=PRO

% verbo
V=V

% verbo auxiliar: tinha, havia, foi. Nao ha etiqueta para verbo auxiliar no Unitex, mas para nao perder essa informacao, cria-se uma
VAUX=VAUX

% simbolo de moeda: R$. Nao ha etiqueta para simbolos de moeda no Unitex, mas para nao perder essa informacao, cria-se uma
CUR=CUR

APÊNDICE B – Regras de desambiguação

B.1 Regras para o inglês

Conteúdo do arquivo regras_desamb_en.txt

PART
-1 [V]

N+PR
0 {ucfirst($_) eq $_}

PFX
1 “-”

B.2 Regras para o português

Conteúdo do arquivo regras_desamb_pt.txt

N+Pr
0 {ucfirst($_) eq $_}

PFX

1 “-”

ANEXO A – Conjunto de etiquetas de MXPOST

Nesse capítulo são apresentados os conjuntos de etiquetas utilizados por MXPOST na etiquetação dos textos em inglês (veja seção A.1) e em português (veja seção A.2).

A.1 Conjunto de etiquetas de MXPOST para o inglês

Nessa seção apresenta-se o conjunto de etiquetas utilizado por MXPOST para a etiquetação dos textos em inglês: o *Penn Tagset* (veja Tabela 9).[1]

Além das etiquetas apresentadas na Tabela 9, esse conjunto de etiquetas possui outras específicas para alguns símbolos de pontuação e caracteres especiais, são elas: #, $, ., ,, :, (,), ‘, “, ’ e ”.

[1] O *Penn Tagset* pode ser consultado em `http://www.mozart-oz.org/mogul/doc/lager/brill-tagger/penn.html`.

Tabela 9: Etiquetas de MXPOST utilizadas para etiquetação de textos em inglês (MARCUS et al., 1993)

Etiqueta	**Descrição**	**Exemplo**
CC	conjunção coordenativa	and
CD	numeral cardinal	third
DT	determinante	the
EX	*there* existencial	there is
FW	palavra estrangeira	d'hoevre
IN	preposição/conjunção subordinativa	in, because
JJ	adjetivo	green
JJR	adjetivo comparativo	greener
JJS	adjetivo superlativo	greenest
LS	item de uma lista	1)
MD	verbo modal	could, will
NN	substantivo singular ou de massa	table
NNS	substantivo plural	tables
NNP	nome próprio singular	John
NNPS	nome próprio plural	Vikings
PDT	pré-determinante	**both** the boys
POS	terminação de possessivo	friend**'s**
PRP	pronome pessoal	I, he, it
PRP$	pronome possessivo	my, his
RB	advérbio	however, usually
RBR	advérbio comparativo	better
RBS	advérbio superlativo	best
RP	partícula	give **up**
SYM	símbolo (matemático or científico)	
TO	*to*	**to** go, **to** him
UH	interjeição	uhhuhhuhh
VB	verbo no infinitivo	take
VBD	verbo no passado	took
VBG	verbo no gerúndio/particípio presente	taking
VBN	verbo no particípio passado	taken
VBP	verbo no presente e não na terceira pessoa do singular	take
VBZ	verbo no presente e na terceira pessoa do singular	takes
WDT	determinante-*wh*	which
WP	pronome-*wh*	who, what
WP$	pronome-*wh* possessivo	whose
WRB	advérbio-*wh*	where, when

A.2 Conjunto de etiquetas de MXPOST para o português

Nessa seção apresenta-se o conjunto de etiquetas (veja Tabela 10) utilizado por MXPOST para a etiquetação dos textos em português (ALUÍSIO et al., 2003).

Tabela 10: Etiquetas de MXPOST utilizadas para etiquetação de textos em português

Etiqueta	Descrição	Exemplo
ADJ	adjetivo	grande, coloridas, interessante
ADV-KS-REL	advérbio relativo subordinativo	onde, quando, como
ADV-KS	advérbio conectivo subordinativo	aonde, como, quando
ADV	advérbio não-subordinado	sempre, bem, ontem, aqui
ART	artigo	o, a, os, as, um, uma, uns, umas
KC	conjunção coordenativa	e, mas, ou, logo
KS	conjunção subordinativa	que, embora, se
IN	interjeição	oi, olá, ai, ah
N	nome	cadeira, povo, livro
NPROP	nome próprio	Maria, Unicamp
NUM	numeral	cem, 2.200
PCP	particípio (ou adjetivo)	assustada, encerrado
PDEN	palavra denotativa	até, somente, então, eis
PREP	preposição	de, para, com, por
PROPESS	pronome pessoal	ela, nos, a, você, comigo
PRO-KS-REL	pronome conectivo subordinativo relativo	que, a qual, quem, cujo
PRO-KS	pronome conectivo subordinativo	quem, quantas, que
PROSUB	pronome substantivo	ninguém, algo, nada
PROADJ	pronome adjetivo	nenhum, cada, vários
VAUX	verbo auxiliar	tinha, havia, foi
V	verbo	moro, gostaria, será, caiu
CUR	símbolo de moeda corrente	R$, US$

Além das etiquetas principais apresentadas na Tabela 10, existem etiquetas complementares usadas, por exemplo, para identificar palavras estrangeiras (veja (ALUÍSIO et al., 2003)).[2]

[2]O guia de etiquetação manual do corpus Mac-Morpho – usado para treinar MXPOST para o português – está disponível em `http://www.nilc.icmc.usp.br/lacioweb/manuais.htm`.

ANEXO B – Conjunto de etiquetas de Unitex

Antes de apresentar os conjuntos de etiquetas usados por Unitex para os idiomas inglês e português vale lembrar, aqui, que a análise de Unitex para uma dada palavra possui o seguinte formato:

palavra,canônica.categoria(+traços)*(:atributos)*.

Nesse formato, a categoria gramatical de **palavra** pode ser dada por apenas uma etiqueta (**categoria**) ou a combinação de vários códigos semânticos e gramaticais (**traços**) separados por um caractere "+". **(:atributos)***, por sua vez, é uma combinação de zero ou mais conjuntos de atributos morfológicos (cada atributo é representado por apenas um caractere) separados pelo caractere ":".

Assim, a Tabela 11 apresenta alguns dos códigos semânticos e gramaticais usados tanto para o inglês quanto para o português. Em seguida, as etiquetas de Unitex utilizadas para a identificação das categorias gramaticas e dos atributos, específicos para cada língua, são apresentados separadamente mas seções B.1 (para o inglês) e B.2 (para o português).

Tabela 11: Alguns códigos de Unitex utilizados para representar gramática e semântica

Etiqueta	**Descrição**	**Exemplos**	
		`pt`	`en`
`z1`	língua usual	piada	joke
`z2`	língua especializada	encafuar	floppy disk
`z3`	língua muito especializada	dzeta	serialization
`Abst`	abstrato	bom gosto	patricide
`Anl`	animal	cavalo	horse
`AnlColl`	animal coletivo	manada	flock
`Conc`	concreto	mesa	chair
`ConcColl`	concreto coletivo	trânsito	rubble
`Hum`	humano	diplomanta	teacher
`HumColl`	humano coletivo	velha guarda	parliament
`t`	verbo transitivo	morder	kill
`i` (`en`)	verbo intransitivo	–	agree
`x` (`pt`)	verbo intransitivo	encalhar	–

B.1 Conjunto de etiquetas de Unitex para o inglês

A maioria das etiquetas usadas por Unitex para o idioma inglês, e apresentadas nesta seção, estão presentes em (PAUMIER, 2006) (e em sua versão em português[1]). Porém, outras etiquetas não citadas em (PAUMIER, 2006) foram derivadas por meio da análise de corpus.

Sendo assim, as Tabelas 12 e 13 apresentam as etiquetas usadas para representar, respectivamente, as categorias gramaticais e os atributos (informações flexionais e traços) das palavras em inglês.

[1] A versão em português dos quatro primeiros capítulos do manual (PAUMIER, 2006) pode ser obtida em: `http://ladl.univ-mlv.fr/brasil/`.

Tabela 12: Etiquetas de Unitex utilizadas para etiquetação das categorias gramaticais em inglês

Etiqueta	**Descrição**	**Exemplos**
A	Adjetivo	fabulous
ADV|ADVA	Advérbio	actually, somewhat
CONJ	Conjunção de coordenação	but
CONJS	Conjunção de subordinação	because
DET	Determinante	each, either
DET+Dnum	Número	eight, hundred
INTJ	Interjeição	eureka
N	Substantivo	table
N+PR	Nome próprio	John, Brazil
PART	Partícula	off, on, to
PFX	Prefixo	re
PRED	Pré-determinante	about, almost, only
PREP	Preposição	without
PRO|PRON	Pronome	you, whatever
V	Verbo	take, eat
V+aux	Verbo auxiliar	could, can, cannot
X|XI	Palavra extrangeira	–

B.2 Conjunto de etiquetas de Unitex para o português

A maioria das etiquetas usadas por Unitex para o idioma português, apresentadas nas Tabelas 14, 15 e 16, estão presentes em (MUNIZ, 2004) enquanto algumas foram obtidas por meio da análise de corpus.

Sendo assim, as Tabelas 14, 15 e 16 apresentam as etiquetas usadas para representar, respectivamente, as categorias gramaticais e os atributos (informações flexionais e traços) das palavras em português.

Tabela 13: Etiquetas de Unitex utilizadas para etiquetação dos atributos em inglês

Etiqueta	**Descrição**	**Exemplos**
`m`	masculino	–
`f`	feminino	–
`n`	neutro	–
`s`	singular	team,.N:s
`p`	plural	researchers,researcher.N+Hum:p
`1,2,3`	1ra, 2da, 3ra pessoa	is,be.V:P3s
`P`	presente do indicativo	prove,.V:W:P1s:P2s:P1p:P2p:P3p
`I`	imperfeito do indicativo	reproduced,reproduce.V:K:I1s:I2s: I3s:I1p:I2p:I3p
`S`	presente do subjuntivo (verbos) superlativo (adjetivos e advérbios)	– worst,ill.A:S
`T`	imperfeito do subjuntivo	–
`Y`	presente do imperativo	–
`C`	presente do condicional (verbos) comparativo (adjetivos e advérbios)	– worse,ill.A:C
`J`	pretérito	–
`W`	infinitivo	study,.V:W:P1s:P2s:P1p:P2p:P3p
`G`	gerúndio (particípio presente)	having,have.V:G
`K`	particípio passado	taken,take.V:K
`F`	futuro	–
`Ddéf|Ddef`	determinante definido	the,.DET+Ddéf:s:p
`Dind`	determinante indefinido	a,.DET+Dind:s
`Ddem`	determinante demonstrativo	this,.DET+Ddem:s
`DetQ`	determinante interrogativo	which,.DET+DetQ
`Dadj`	determinante adjetival	all,.DET+Dadj
`RelQ`	pronome interrogativo	which,.PRO+RelQ:s:p
`Refl`	pronome reflexivo	itself,.PRO+Refl:3ns
`Pdem`	pronome demonstrativo	this,.PRO+Pdem:s
`Poss`	pronome possessivo determinante possessivo	its,.PRO+Poss3ns:s:p its,.DET+Poss3ns:s:p

Tabela 14: Etiquetas de Unitex utilizadas para etiquetação das categorias gramaticais em português

Etiqueta	**Descrição**	**Exemplos**
A	Adjetivo	bonito, bonitas, aprazível, simples
ABREV	Abreviatura	ml, mm
ADV	Advérbio	abaixo, misericordiosissimamente, mesmo
CONJ	Conjunção de coordenação	mas, mais, mal
DET+Art	Artigo	o, umas
DET+Num	Numeral	segundo, duplo
INTERJ	Interjeição	ah, ih, olá, oi
N	Substantivo	menino, meninão, lápis, ajuda
N+Pr	Nome próprio	João, Silva
PFX	Prefixo	super, pós, sub
PREP	Preposição	ante, de
PRO	Pronome	senhora, eu
SIGL	Sigla	ONU, PDT, OTAN, USP
V	Verbo	cantaríamos, cantarias, cantaria

Tabela 15: Etiquetas de Unitex utilizadas para etiquetação dos atributos em português

Etiqueta	**Descrição**	**Exemplos**
m	masculino	menino,.N:ms
f	feminino	ajuda,.N:fs
s	singular	menino,.N:ms
p	plural	meninos,menino.N:mp
1,2,3	1ra, 2da, 3ra pessoa	eu,.PRO+Pes:N1ms:N1fs
P	presente do indicativo	há,haver.V:P3s:Y2s
I	pretérito imperfeito do indicativo	estava,estar.V:I1s:I3s
S	presente do subjuntivo (verbos) superlativo (adjetivos)	cobre,cobrar.V:S1s:S3s:Y3s amabilíssimo,amável.A:Sms
T	imperfeito do subjuntivo	cobrasse,cobrar.V:T1s:T3s
Y	imperativo	pára,parar.V:P3s:Y2s
C	futuro do pretérito (verbos) cardinal (numerais)	cantarias,cantar.V:C2s milhões,milhão.DET+Num:Cmp
J	pretérito perfeito do indicativo	recebeu,receber.V:J3s
W	infinitivo	provar,.V:W1s:W3s:U1s:U3s
G	gerúndio	disparando,disparar.V:G
K	particípio	feita,fazer.V:K
F	futuro do presente do indicativo (verbos) fracionário (numerais)	fará,fazer.V:F3s quarto,.DET+Num:Fms
Q	pretérito mais que perfeito do indicativo	reproduziram,reproduzir.V:J3p:Q3p
U	futuro do subjuntivo	testar,.V:W1s:W3s:U1s:U3s
A	aumentativo (substantivos e adjetivos) forma acusativa (pronomes)	meninão,menino.N:Ams os,ele.PRO+Pes:A3mp
D	diminutivo (substantivos e adjetivos) forma dativa (pronomes)	menininha,menino.N:Dfs no,ele.PRO+Pes:O3ms:A3ms:D3ms
N	forma nominativa (pronomes)	eu,.PRO+Pes:N1ms:N1fs
O	forma oblíqua (pronomes) ordinal (numerais)	na,ele.PRO+Pes:O3fs:A3fs:D3fs segundo,segundo.DET+Num:Oms
R	forma reflexa (pronomes)	se,ele.PRO+Pes:R3ms:R3fs:R3mp:R3fp

Tabela 16: Etiquetas de Unitex utilizadas para etiquetação dos atributos em português (cont.)

Etiqueta	**Descrição**	**Exemplos**
`M`	multiplicativo (numerais)	duplo,duplo.DET+Num:Mms
`L`	coletivo (numerais)	–
`Def`	artigo definido	o,o.DET+Art+Def:ms
`Ind`	indefinido (artigos e pronomes)	umas,um.DET+Art+Ind:fp
`Dem`	pronome demonstrativo	este,.PRO+Dem:ms
`Rel`	pronome relativo	quando,.PRO+Rel:ms:mp:fs:fp
`Int`	pronome interrogativo	que,.PRO+Int:ms:mp:fs:fp
`Tra`	pronome de tratamento	senhora,senhor.PRO+Tra:3fs
`Pos`	pronome possessivo	meu,eu.PRO+Pos:1ms
`Pes`	pronome pessoal	eu,eu.PRO+Pes:N1ms:N1fs