CURUPIRA: UM PARSER FUNCIONAL PARA A LÍNGUA PORTUGUESA[1]

Ronaldo Teixeira Martins; Ricardo Hasegawa; Maria das Graças Volpe Nunes
Núcleo Interinstitucional de Lingüística Computacional - NILC

1. Introdução

CURUPIRA, do tupi CURU (menino) e PIRA (corpo), também conhecido como CAAPORA (ou habitante do mato), é um mito que integra o folclore amazônico, geralmente representado como um menino de pele escura, cabelos cor-de-fogo, pés virados para trás (para poder despistar os caçadores) e portador do dom da invisibilidade. É muitas vezes representado montando um cateto (espécie de porco-do-mato) e carregando um pequeno machado feito de casco de jabuti. Filho de Tupã, mas criado por Anang, Curupira, ou mãe do mato, defende a floresta de seus invasores, sendo responsável por toda má sorte que atinge os caçadores, iludindo-os, fazendo que se percam e andem em círculos, que persigam caças imaginárias até a morte, ou que suas armas sempre errem o alvo, entre outras malandragens.

CURUPIRA é também o nome do analisador automático de sentenças da língua portuguesa desenvolvido pelo NILC. A escolha do nome se deveu, principalmente, a três características do *parser* ora consideradas fundamentais. Em primeiro lugar, trata-se de um analisador sintático que se pretende, como a lenda, permanentemente jovem, em contínuo processo de reconstrução, dada a complexidade dos problemas relativos à etiquetação sintática automática das sentenças da língua portuguesa (como, de resto, de qualquer língua natural). Em segundo lugar, é um analisador sintático que pretende respeitar e preservar a riqueza estrutural das sentenças da língua portuguesa, entendidas (e representadas) como árvores de que o texto (na verdade uma floresta sintática) seria composto. O *parser*, diferentemente de outras ferramentas à sua semelhança, não faz uso de estratégias de poda, de simplificação, ou de redução, procurando etiquetar, sintaticamente, cada um dos itens lexicais da sentença, por menores e menos expressivos que sejam. Procura preservar, desta forma, todas as variedades estruturais das sentenças da língua, razão pela qual, aliás, faz uso de um conjunto razoavelmente extenso de etiquetas sintáticas, que acompanham, quase sempre, e com granularidade freqüentemente superior, os rótulos previstos pela Nomenclatura Gramatical Brasileira (NGB). Finalmente, o analisador sintático, que opera em sentido descendente recursivo, da esquerda para a direita, perde-se, com freqüência, entre as pegadas lexicais deixadas pelo usuário (o verdadeiro Curupira), voltando com freqüência ao ponto de partida por meio de estratégias de *backtracking*.

O CURUPIRA é resultado do processo de desenvolvimento e de isolamento do *parser* que, desde 1997, vem operando (invisivelmente) no interior do ReGra - o revisor gramatical automático para a língua portuguesa também desenvolvido pelo NILC. Como no caso da rotina que integra o ReGra, o CURUPIRA toma, como premissa, a hipótese de que as sentenças da língua portuguesa (como as de todas as línguas naturais) podem ser representadas por estruturas de tipo arbóreo, e que essa representação é não apenas útil mas imprescindível para a determinação das relações de dependência que se observam entre os itens lexicais co-ocorrentes. O *parser* concebe o texto, pois, como uma floresta repleta de árvores, cada uma das quais definível e explorável em separado, sem que a consideração da parte torne obrigatória a observação de sua relação com o todo que integra. Por este

[1] Trabalho realizado com o apoio da FAPESP, ITAUTEC-PHILCO S.A. e CNPq.

motivo, admite, como dado de entrada, apenas sentenças isoladas, assim entendidas as seqüências de itens lexicais delimitadas por um conjunto de marcadores específicos (como o ponto e o marcador de fim de parágrafo, por exemplo). Conseqüentemente, o CURUPIRA, embora não faça qualquer restrição quanto à dimensão, quanto à forma, ou quanto ao domínio da sentença de entrada, não é capaz de resgatar relações de natureza extra-sentencial, seja por referência ao co-texto imediato, seja por referência ao contexto situacional, incluídas as substituições, as elipses, as pronominalizações e todas as relações anafóricas. Cabe observar ainda que, diferentemente do que ocorre em relação ao ReGra, o CURUPIRA supõe que a sentença de entrada esteja correta e provê, não apenas uma, mas todas as possibilidades sintáticas previstas pela gramática que lhe serve de base.

Cabe assinalar ainda que o CURUPIRA é um *parser* simbólico e funcional. É um *parser* simbólico, na medida em que faz uso de um conjunto explícito de regras de manipulação de palavras e classes de palavras, concebido pela intuição do falante a partir do exame de um fragmento de corpus, e adaptado, historicamente, aos vários contra-exemplos que a experiência de análise automática foi fornecendo. Nesse sentido, o CURUPIRA não faz uso de estratégias subsimbólicas, estatísticas ou extensivamente baseadas em *corpora*. Há uma indicação de prioridade de aplicação em relação às regras que compõem a gramática, e há também uma indicação de prioridade de classificação para as entradas que constituem o dicionário, mas uma e outra foram obtidas antes por critérios relativos à conveniência da análise (a substantivação de adjetivos é mais simples que a adjetivação de substantivos, por exemplo), do que por uma confirmação empírica de sua freqüência de ocorrência no recorte de *corpus* analisado. As únicas exceções dizem respeito às entradas e às regras que, embora incluídas no dicionário e na gramática, caracterizam já um uso muito pouco freqüente (ou antes geograficamente muito localizado) do português brasileiro (caso das formas verbais da segunda pessoa do plural, por exemplo).

Por fim, o CURUPIRA é um *parser* sobretudo funcional porque, ao fazer uso de um conjunto de etiquetas sintáticas muito detalhado, informa não apenas a estrutura sintagmática da sentença, mas suas relações (funcionais) de dependência sintática. O vocabulário não-terminal que caracteriza as regras que compõem a gramática sobre a qual opera o CURUPIRA traz não apenas a informação categorial relativa ao núcleo do sintagma (nominal, verbal, etc.), mas a função sintática a que se aplica (sujeito, predicado verbo-nominal, adjunto adverbial oracional, etc.). Os papéis sintáticos são representados, portanto, não apenas por meio de critérios distribucionais (ou posicionais), mas estão contidos na própria estrutura do sintagma (na medida em que se distinguem, por exemplo, as estruturações do sintagma nominal sujeito das estruturas do sintagma nominal objeto direto e de todos os outros sintagmas nominais).

Neste relatório são descritos os objetivos (segunda seção) e a metodologia (terceira seção) utilizada para o desenvolvimento do CURUPIRA. A consideração dos resultados - bem como a avaliação do desempenho da ferramenta - será feita em relatório à parte, dado o estágio ainda preliminar dos testes. A título de anexo, são apresentados os vocabulários não-terminal e terminal e o conjunto de regras da gramática que serve de base ao *parser*.

2. Objetivos

O CURUPIRA é um *parser* autônomo em relação à aplicação. O principal objetivo do CURUPIRA é fornecer, para uma dada sentença de entrada, pertencente ao registro da escrita da variedade culta do português brasileiro, em seu uso referencial[2], todas as suas possibilidades de análise sintática, assim entendidas as possibilidades combinatórias (de hierarquização) das classes de palavras contidas na sentença, independentemente de seu significado. Este objetivo está, por um lado, delimitado por uma determinada concepção de sentença (e de unidade de processamento lingüístico) e de representação da estrutura da sentença; por outro lado, o objetivo não está associado a nenhum tipo de aplicação específica, podendo os resultados do CURUPIRA ser utilizados por toda sorte de ferramenta que demande alguma informação relativa às estruturas sintáticas do português.

Em relação à concepção de sentença, cabe dizer que o CURUPIRA interpreta como "sentença" toda e qualquer seqüência de caracteres isolada por dois delimitadores, representados por um conjunto de caracteres especiais de pontuação ou de comando. São eles: o ponto, o ponto-e-vírgula, os dois-pontos, as reticências, o ponto-de-interrogação, o ponto-de-exclamação e o travessão, entre os sinais de pontuação; e os marcadores de início-de-linha, de fim-de-linha, de início-de-parágrafo, de fim-de-parágrafo, de início-de-coluna, de fim-de-coluna, de início-de-página e de fim-de-página, como caracteres de comando. O tamanho ou a variedade do intervalo entre dois delimitadores não importa ao CURUPIRA, que está preparado para a análise de sentenças de qualquer comprimento. Não há, portanto, restrições quanto ao número de palavras e/ou de caracteres da sentença de entrada. Mas é preciso observar que sentenças muito longas degradam o desempenho da ferramenta, que poderá travar em ambientes computacionais menos robustos. Nenhum gerenciamento da memória alocada é feito pela ferramenta, que procura fazer que a seqüência de caracteres corresponda a pelo menos uma das estruturas sintáticas previstas pela gramática que lhe serve de base. Quando isso não é possível - seja por falha do conjunto de possibilidades sintáticas da ferramenta, seja por omissão do dicionário, seja por efetiva agramaticalidade da sentença de entrada - a ferramenta emite uma mensagem informando ao usuário que o casamento [com alguma das estruturas previstas] não pôde ser realizado. Durante esse processo de estruturação da *string* de entrada, todos os caracteres são considerados significativos, inclusive os espaços em branco, as vírgulas e os outros caracteres especiais que não figuram como delimitadores de sentença.

Dado o foco na sentença como objeto de análise lingüística, o CURUPIRA não fornece nenhuma informação sobre relações de natureza extra-sentencial, ainda que possam ser facilmente resgatáveis por referência ao co-texto mínimo imediato. Dessa forma, o preenchimento de elipses e a recuperação de anáforas, ainda que pudessem ser equacionados pela análise (ou pelo armazenamento) das informações da sentença imediatamente anterior, não constituem objetivos do CURUPIRA, que pretende, tão somente, estruturar a seqüência de caracteres de entrada tal como ela se apresenta, sem promover qualquer tipo de transformação ou de movimento nessa estrutura superficial, por mais pertinentes e importantes que tais modificações possam parecer.

Do ponto de vista da representação da estrutura da sentença, o CURUPIRA optou por uma representação mais funcional do que formal, aproximando-se da notação proposta pela Nomenclatura Gramatical Brasileira (NGB - Portaria MEC 36, de 28/01/59), apesar de suas muitas inconsistências. Essa notação se revelou pertinente, em primeiro lugar, por motivos históricos, dado que o *parser* é produto do isolamento de um analisador sintático que servia de base a uma ferramenta de correção

[2] Aqui em oposição aos outros possíveis usos da linguagem: poético, metalingüístico, fático, conativo, emotivo, nomeadamente, segundo [JAK95].

gramatical e revisão de estilo para a língua portuguesa - o ReGra, que verifica a adequação das sentenças do usuário às estruturas prescritas pelas gramáticas normativas do português, e que se vale, portanto, da nomenclatura em que essas regras estão vertidas. Adicionalmente, o uso de uma notação mais próxima da gramática tradicional observa também a necessidade de fornecer uma representação mais detalhada, incorporando algumas distinções que normalmente são opacificadas pelas notações mais formais (como a distinção entre adjunto adnominal e complemento nominal, por exemplo; ou entre adjunto adnominal e aposto). Na verdade, o conjunto de etiquetas proposto pela NGB foi alterado, em vários momentos, para que pudesse contemplar diferenças estruturais não alcançadas pela simplificação empreendida pela própria NGB, como a diferença entre o adjunto adnominal à esquerda e o adjunto adnominal à direita, ou entre adjunto adverbial oracional e adjunto adverbial local. Por este motivo, o CURUPIRA faz uso de um estoque de etiquetas sintáticas bastante mais numeroso do que a das ferramentas congêneres, sendo este um de seus traços distintivos. O conjunto completo das etiquetas utilizadas na versão 1.0 é apresentado no Anexo I deste relatório.

Por fim, cabe salientar que, entre os objetivos do CURUPIRA, não está o de desambiguar a estrutura sintática das sentenças da língua portuguesa. A ferramenta fornece todas as possibilidades de análise sintática, apresentadas segundo a prioridade de aplicação das regras da gramática que lhe serve de base, sem qualquer compromisso com a indicação da estrutura "correta", ou "mais adequada", ou mesmo "mais provável" para a sentença de entrada. Como o CURUPIRA não realiza nenhuma espécie de análise semântica, tudo o que consegue fazer é combinar palavras e classes de palavras de forma a oferecer, como saída, suas possibilidades de estruturação sintática, muitas das quais se revelarão, evidentemente, descabidas e impertinentes, por não admitirem nenhuma possibilidade de projeção semântica.

3. Metodologia

A estrutura do CURUPIRA é apresentada na Figura 1. Na primeira subseção - dos Recursos, cada um dos módulos desta arquitetura será analisado separadamente; na segunda subseção - dos Procedimentos, será apresentada a interação entre os vários módulos.

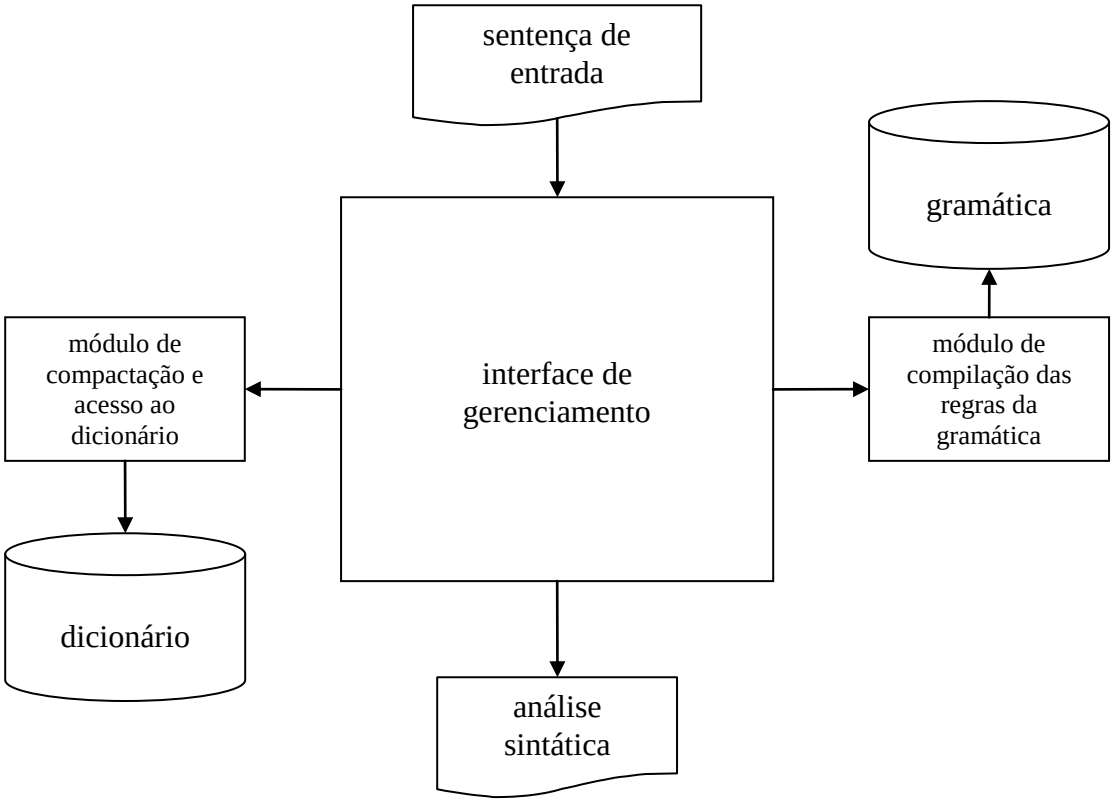


Figura 1. Arquitetura Geral do CURUPIRA

3.1 Recursos

3.1.1. Dicionário

O CURUPIRA opera sobre o mesmo léxico que serve ao ReGra. Trata-se de uma base de dados lexical robusta, não restrita em relação a nenhum domínio, onde estão representadas todas e apenas as lexias simples e compostas da língua portuguesa. Não estão representados morfemas, sejam eles lexicais ou gramaticais, nem expressões complexas que contenham espaços em branco. Este léxico soma hoje cerca de 1,5 milhão de entradas, em formato texto plano, e foi obtido, primeiramente, a partir de varredura automática do corpus do NILC, de cerca de 40 milhões de palavras, compreendendo textos de gênero predominantemente jornalístico (embora estejam também presentes, em menor grau, textos literários, textos técnicos e mesmo redações escolares). A esse processo de composição inicial, em que as entradas foram classificadas manualmente, foram acrescidas outras entradas, por flexão ou derivação automática (caso das formas verbais e dos advérbios de modo, por exemplo), além de outras palavras, obtidas nas várias sessões de testes a que o dicionário foi submetido. As entradas do dicionário possuem o seguinte formato:

cantar=<V.[BI.INT.TD.][FUT-SUBJ.ELE.FUT-SUBJ.EU.INF-PESS.ELE.INF-PESS.EU]N.[a][cantar]0.>

Como se pode verificar, o léxico traz apenas informações de natureza morfossintática. Não há, portanto, nenhuma possibilidade de processamento semântico das sentenças de entrada. Em cada entrada, estão representados: a classe das palavras (substantivo, adjetivo, numeral, nome próprio, abreviatura, sigla, prefixo, interjeição, conjunção, preposição, artigo, advérbio, verbo, pronome), as subclassificações pertinentes a cada classe (pronome pessoal, possessivo, demonstrativo, indefinido, interrogativo, de tratamento, em relação aos pronomes, por exemplo; ou numeral multiplicativo, cardinal, ordinal e fracionário, em relação aos numerais); o gênero, sempre que pertinente (masculino, feminino, ou uniforme); o número, sempre que pertinente (singular, plural ou invariável); o tempo e o modo, para as formas verbais (presente do indicativo, futuro do subjuntivo, etc.); o número e a pessoa, para as formas verbais e para os pronomes pessoais (primeira pessoa do singular, segunda pessoa do plural, etc.); o grau, sempre que pertinente (positivo, aumentativo, diminutivo); a transitividade, para os verbos (transitivo direto, intransitivo, etc.); a regência, para os substantivos, adjetivos, verbos e advérbios que exigem complemento preposicionado; o tipo, para os advérbios (de tempo, de modo, de intensidade, etc.); e a forma canônica (ou a forma de citação da palavra, pela qual ela está relacionada às outras entradas do dicionário), para todas as formas.

Na medida em que nenhuma informação de natureza semântica está representada no dicionário, não são registradas diferenças entre acepções distintas de uma mesma entrada, nem mesmo quando correspondem a formas canônicas diferentes ou a diferentes classes gramaticais. Nos casos de homonímia, as diferentes classificações possíveis para cada verbete são ordenadas, na mesma entrada, segundo a conveniência de análise, que está normalmente amparada na freqüência de ocorrência para o falante nativo da língua. A desambiguação categorial é feita a) pela freqüência de ocorrência, indicada no próprio dicionário, b) pelo próprio conjunto de possibilidades sintáticas previstas pela gramática, ou c) por um conjunto (bastante restrito) de regras de desambiguação incorporadas à ferramenta.

Informações mais detalhadas sobre o dicionário podem ser obtidas em [NUN96]. O conjunto dos traços categoriais utilizado no dicionário, e mobilizado pelas regras da gramática (a título de vocabulário terminal), é apresentado no Anexo II deste relatório.

A par do léxico, o CURUPIRA também é integrado por um conjunto de lexias complexas (locuções prepositivas, locuções conjuncionais, locuções e expressões latinas, etc.) que, por força de restrições dos algoritmos de busca e compactação, não foi possível incluir no léxico. Este conjunto de expressões (que contêm espaços em branco) está dicionarizado dentro do próprio algoritmo de análise sintática e caracteriza antes parte da gramática.

3.1.2. Módulo de compactação e acesso ao dicionário (KLS)

Kowaltowski, Lucchesi e Stolfi [KOW93; KOW95a; KOW95b; KOW98], pesquisadores do IC-Unicamp, desenvolveram um trabalho sobre o uso de autômato finito para representação de grandes vocabulários com o intuito de desenvolver um aconselhador ortográfico para a língua portuguesa. O sistema resultante, conhecido como KLS (iniciais de Kowaltowski, Lucchesi e Stolfi), permite que a representação do léxico do NILC, com cerca de 1,5 milhão de entradas, mais os atributos sintáticos, seja codificado em um autômato que ocupa menos de 1,3 Mb. Isto se deve, principalmente, ao grande número de formas derivadas na língua portuguesa. Por exemplo, um verbo regular do português

apresenta 51 flexões distintas, cujos sufixos podem ser compartilhados entre todos os outros verbos regulares.

O algoritmo de busca para uma dada palavra também é muito eficiente. Vale lembrar que o acesso à estrutura do autômato é feito de forma direta, sendo necessárias apenas *n.m* operações de comparação, onde *n* é o número de letras da palavra sendo procurada, e *m* é o número médio de transições por estado. Em outras palavras, dado o estado inicial, o algoritmo percorre o autômato utilizando letras consecutivas da palavra para selecionar as transições, até que um estado final seja alcançado ou não existam mais transições válidas.

3.1.3. Gramática

A gramática que serve ao CURUPIRA pode ser definida pela quíntupla <FRASE, V, t, R, P>, em que FRASE é o símbolo inicial (qualquer intervalo entre dois delimitadores de sentença); V corresponde ao vocabulário não-terminal (composto por um conjunto de etiquetas sintáticas que, como já afirmado anteriormente, procura se aproximar da notação proposta pela NGB); t corresponde ao vocabulário terminal, ou conjunto de traços categoriais presentes no dicionário (como descrito na seção 3.1.1); R é um conjunto de aproximadamente 600 regras de reescrita categorial, conforme sintaxe definida a seguir; e P é a prioridade de aplicação das regras de reescrita, atribuída ora pela freqüência de ocorrência, ora pela conveniência de análise. O conjunto dos símbolos não-terminais é apresentado no Anexo I deste relatório. O conjunto das regras de reescrita categorial, com a indicação de sua prioridade de aplicação, compõe o Anexo III.

A gramática corresponde a um arquivo texto plano que, antes de ser submetido ao CURUPIRA, passa pelo processo de compilação descrito na próxima seção. A sintaxe das regras de reescrita categorial é a que se segue:

A#B#C

em que:
A = símbolo
B = prioridade de aplicação da regra: número inteiro igual ou superior a 1 (1 = prioridade máxima)
C = regras de reescrita (conjunto de possibilidades sintáticas de realização do símbolo expresso por A)

+ = justaposição, com separação por espaço em branco (apenas)
[X] = opcionalidade (x é opcional)
X(Y) = Y deve ser atributo de X
(i) = indexado (os termos portadores de (i) devem concordar em número, pessoa e gênero)
{X,Y} = exclusividade: ou X, ou Y
'entre aspas simples' = entradas do dicionário
<X> = forma canônica

Um exemplo de regra de reescrita categorial é apresentado a seguir:

```
SUJ_SIMPLES#5#AADNE + [APOSTO] + nucleo(SUJ2) + [AADND] + [APOSUJ]
```

O CURUPIRA foi especialmente desenvolvido para o processamento de sentenças que pertencem ao registro da escrita da variedade culta do português brasileiro, particularmente das estruturas sintáticas mais diretas, simples, características de textos informativos e argumentativos. Em função desse escopo, na composição da gramática que serve de base ao CURUPIRA, as regras de reescrita categorial não prevêem, normalmente, inversões sintáticas muito radicais, topicalizações, clivagens, anacolutos e outras rupturas da estruturação sintática convencional do registro da escrita do português brasileiro. No entanto, e a despeito de o CURUPIRA tomar, como ponto de partida, apenas sentenças gramaticais da língua portuguesa, foram mantidas, na especificação da gramática, algumas possibilidades sintáticas que, embora possam ser consideradas agramaticais pelas gramáticas prescritivas, são freqüentemente praticadas por usuários do português. No vocabulário não-terminal da gramática estão previstas, portanto, etiquetas que correspondem a desvios freqüentes do usuário, principalmente na estruturação do objeto direto e do objeto indireto.

3.1.4. Compilador

O compilador é um programa implementado em Visual C++ .Net para a plataforma Windows que converte um arquivo texto plano contendo as regras de reescrita categorial em vários outros arquivos contendo classes de objetos em linguagem de programação C++. Uma classe de objetos é formada pela união de todas as regras de reescrita categorial que possuam o mesmo símbolo expresso por A (sintaxe na seção 3.1.3). As regras de reescrita expressas por C são convertidas em métodos da classe e controladas por um gerenciador de regras que aplica a prioridade dada por B. Este gerenciador de regras também é responsável por parte do controle do mecanismo de *backtracking* e é responsável por aplicar alguns mecanismos de otimização. Por exemplo, se após o gerenciador aplicar uma determinada regra em determinada posição da frase e a regra não for válida para esta posição, ele irá desabilitar esta regra para esta posição da frase. Este mecanismo permite que regras não candidatas sejam disparadas apenas uma única vez por análise.

Uma das características aproveitadas pelo compilador é o mecanismo de herança. Assim, a classe-base inclui os métodos que todas as classes derivadas subseqüentes devem possuir em comum (ou seja, quase todos os métodos com os mecanismos de gerenciamento das regras); e as classes derivadas incluem, principalmente, os métodos que correspondem às regras de reescrita categorial. A estrutura hierárquica é apresentada a seguir:

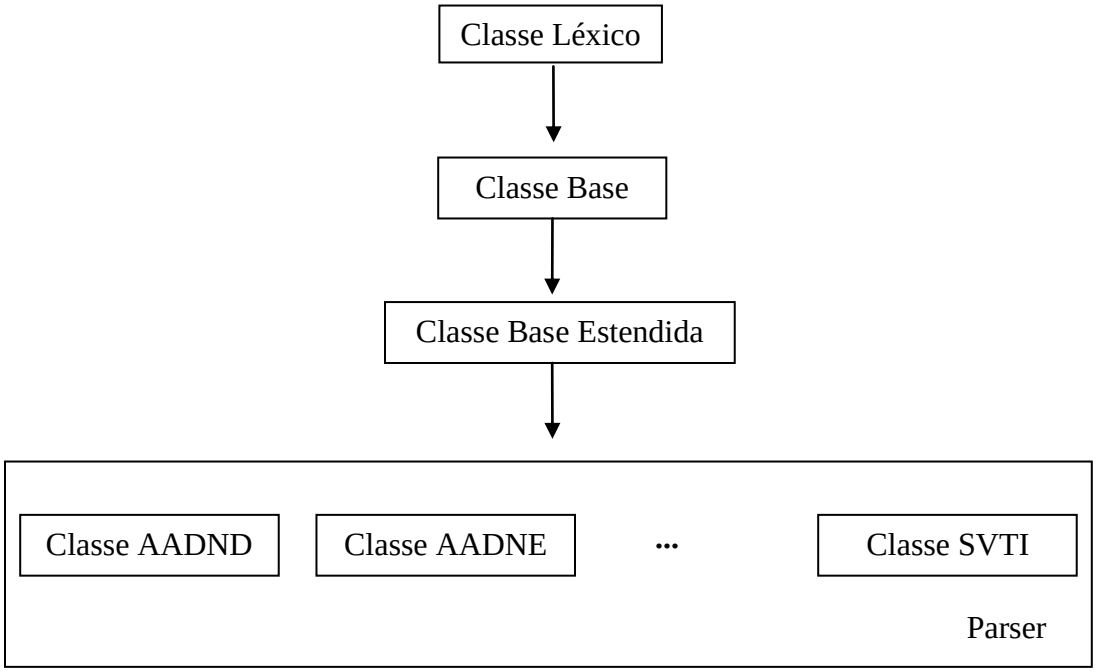


Figura 2: Estrutura hierárquica das classes-base e classes derivadas

A classe **Léxico** é uma interface responsável pela interação entre o dicionário com informações lexicais e as regras de reescrita categorial. A classe **Base** contém os principais mecanismos de gerenciamento das regras de reescrita categorial. Finalmente, a classe **Base Estendida** é criada pelo compilador e deve conter as peculiaridades do *parser*. Por exemplo, as regras com vocábulos terminais não são separadas em classes distintas de objetos, mas acrescentadas como métodos da classe Base Estendida.

Todo o código fonte gerado pelo compilador, mais o código fonte da arquitetura básica herdado do ReGra, são incorporados em um projeto em Visual C++ que ao ser executado gera uma DLL (Dynamic Link Library). Posteriormente, esta DLL mais o dicionário compactado podem ser carregados por um programa executável que contém uma interface para chamar o *parser* e visualizar o resultado.

3.1.5. Interface

O CURUPIRA interpreta como "sentença" qualquer intervalo de palavras entre um conjunto especificado de delimitadores sentenciais, a saber: o ponto, o ponto-e-vírgula, os dois-pontos, as reticências, o ponto-de-interrogação, o ponto-de-exclamação e o travessão, entre os sinais de pontuação; e os marcadores de início-de-linha, de fim-de-linha, de início-de-parágrafo, de fim-de-parágrafo, de início-de-coluna, de fim-de-coluna, de início-de-página e de fim-de-página, como caracteres de comando; e quaisquer outros caracteres não previstos no alfabeto da ferramenta. Os textos de entrada devem estar vertidos no conjunto de caracteres alfanuméricos do padrão ANSI, admitida também a utilização de alguns caracteres textuais mais comuns (vírgula, hífen, maior-que, menor-que, etc.). O formato da saída do CURUPIRA é caracterizado por uma estrutura arbórea, representada de forma indentada, baseada no formalismo da parentetização etiquetada. As etiquetas sintáticas são assinaladas

aos sintagmas, em seus vários níveis, com a indicação da função sintática que as palavras e as expressões assumiriam no contexto da sentença. O repertório das funções sintáticas, por força do ReGra, aplicativo do qual a ferramenta provém, acompanha o conjunto de rótulos previstos pela Nomenclatura Gramatical Brasileira, com algumas modificações. A lista dos símbolos (terminais e não-terminais) utilizados pelo CURUPIRA é apresentada anexa a estas instruções.

A utilização do CURUPIRA é feita por mieo da interface gráfica desenhada para permitir de forma clara a obtenção de toda informação da análise realizada pelas regras do *parser* (Figura 3). Os principais elementos da interface gráfica são descritos a seguir.

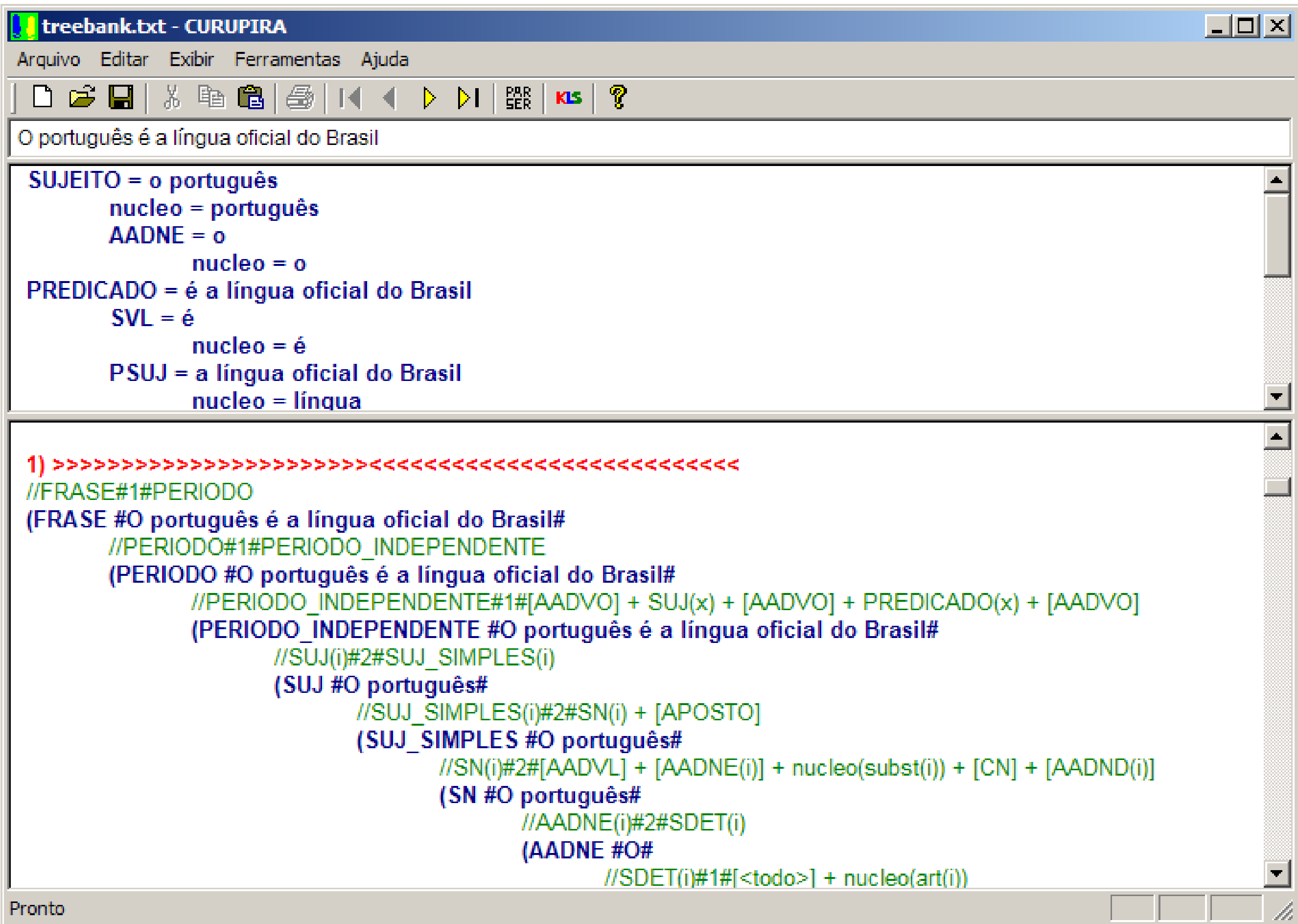


Figura 3: Interface gráfica do CURUPIRA.

Barra de menus - Nesta área ficam posicionados todos os menus que compõem a interface. Todos os principais recursos podem ser acessados através destes menus.

Barra de ferramentas - Esta área contém os botões com ícones representativos das funções do CURUPIRA. Trata-se de um tipo de atalho aos recursos por meio de linguagem gráfica, e não por textos como se faz na barra de menus.

Janela de edição - Trata-se da caixa de entrada de texto localizada na primeira janela. Permite a edição da frase a ser analisada.

Janela de comentários - Nesta área podem ser feitas anotações.

Janela de visualização - Permite visualizar a saída fornecida pelo CURUPIRA de duas maneiras, dependendo do objetivo da pesquisa ou simplesmente da preferência do usuário.

- O primeiro modo apresenta somente os itens lexicais e suas categorias e a(s) árvore(s) sintática(s) que casaram (ou seja, que encontraram alguma estrutura sintática equivalente).

- O segundo modo acrescenta todos os atributos dos itens lexicais e as regras de reescrita categorial que foram usadas durante a análise sintática.

Para a frase "O português é a língua oficial do Brasil" podemos visualizar no segundo modo o seguinte resultado:

```
ART/PROINDVAR/PROPOS/PREP/CONTR/NUM + *ADJ/PART* + VERBO/~SUBST --> SUBST (português)
0) O --> ART
//  Art(masculino, singular, artigo definido) Pro(masculino, singular, [terceira pessoa, numero
indefinido]) Sub(masculino, singular, grau nulo)
1) português --> SUBST
//  Adj(masculino, singular, grau nulo) Nom(masculino, singular, grau nulo)
2) é --> VERBO[ser]-> Lig TransInd
//  Ver(presente, terceira pessoa, singular, genero indefinido)
3) a --> ART
//  Art(feminino, singular, artigo definido) Pre() Pro(feminino, singular, [terceira pessoa, numero
indefinido]) Abrev(masculino, singular) Sub(masculino, singular, grau nulo)
4) língua --> SUBST
//  Sub(masculino, singular, grau nulo) Sub(feminino, singular, grau nulo)
5) oficial --> ADJ
//  Adj(masculino, singular, grau nulo) Adj(feminino, singular, grau nulo) Sub(masculino, singular,
grau nulo) Sub(feminino, singular, grau nulo)
6) do --> PREPOSICAO
//  Pre()
7) do --> ART
//  Art(masculino, singular, artigo definido)
8) Brasil --> NOMPRO
//  Nom(masculino, singular, grau nulo)
9) #>>>>>>>>>> DELIMITADOR
Número de itens: 9


1) >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

//FRASE#1#PERIODO
(FRASE #O português é a língua oficial do Brasil#
   //PERIODO#1#PERIODO_INDEPENDENTE
   (PERIODO #O português é a língua oficial do Brasil#
      //PERIODO_INDEPENDENTE#1#[AADVO] + SUJ(x) + [AADVO] + PREDICADO(x) + [AADVO]
      (PERIODO_INDEPENDENTE #O português é a língua oficial do Brasil#
         //SUJ(i)#2#SUJ_SIMPLES(i)
         (SUJ #O português#
            //SUJ_SIMPLES(i)#2#SN(i) + [APOSTO]
            (SUJ_SIMPLES #O português#
               //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
               (SN #O português#
                  //AADNE(i)#2#SDET(i)
                  (AADNE #O#
                     //SDET(i)#1#[<todo>] + nucleo(art(i))
                     (SDET #O#
                        (nucleo #o# artigo)
                     )
                  )
                  (nucleo #português# subst)
               )
            )
         )
         //PREDICADO(i)#1#PREDVN(i)
         (PREDICADO #é a língua oficial do Brasil#
            //PREDVN(i)#3#SVTI(i) + OI(x) + POBJ(x)
            (PREDVN #é a língua oficial do Brasil#
               //SVTI(i)#1#[verbo(aux) + [verbo(aux)]] + verbo(vti(i))
               (SVTI #é#
```

```
                  (nucleo #é# verbo)
               )
               //OI(i)#4#OI_SIMPLES(i)
               (OI #a língua#
                  //OI_SIMPLES(i)#1#[poi] + SN(i)
                  (OI_SIMPLES #a língua#
                     //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                     (SN #a língua#
                        //AADNE(i)#2#SDET(i)
                        (AADNE #a#
                           //SDET(i)#1#[<todo>] + nucleo(art(i))
                           (SDET #a#
                              (nucleo #a# artigo)
                           )
                        )
                        (nucleo #língua# subst)
                     )
                  )
               )
               //POBJ(i)#1#SADJ(i)
               (POBJ #oficial do Brasil#
                  //SADJ(i)#1#[AADVL] + nucleo(adj(i)) + [CN] + [SADJ(i)]
                  (SADJ #oficial do Brasil#
                     (nucleo #oficial# adj)
                     //CN#2#SP
                     (CN #do Brasil#
                        //SP#1#preposicao + SN
                        (SP #do Brasil#
                           (nucleo #de# preposicao)
                           //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                           (SN #do Brasil#
                              //AADNE(i)#2#SDET(i)
                              (AADNE #do#
                                 //SDET(i)#1#[<todo>] + nucleo(art(i))
                                 (SDET #do#
                                    (nucleo #o# artigo)
                                 )
                              )
                              (nucleo #brasil# nome proprio)
                           )
                        )
                     )
                  )
               )
            )
         )
      )
   )
)


2) >>>>>>>>>>>>>>>>>>>>>>>><<<<<<<<<<<<<<<<<<<<<<<<<<<

//FRASE#1#PERIODO
(FRASE #O português é a língua oficial do Brasil#
   //PERIODO#1#PERIODO_INDEPENDENTE
   (PERIODO #O português é a língua oficial do Brasil#
      //PERIODO_INDEPENDENTE#1#[AADVO] + SUJ(x) + [AADVO] + PREDICADO(x) + [AADVO]
      (PERIODO_INDEPENDENTE #O português é a língua oficial do Brasil#
         //SUJ(i)#2#SUJ_SIMPLES(i)
         (SUJ #O português#
            //SUJ_SIMPLES(i)#2#SN(i) + [APOSTO]
            (SUJ_SIMPLES #O português#
               //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
               (SN #O português#
                  //AADNE(i)#2#SDET(i)
                  (AADNE #O#
                     //SDET(i)#1#[<todo>] + nucleo(art(i))
                     (SDET #O#
                        (nucleo #o# artigo)
                     )
                  )
                  (nucleo #português# subst)
               )
```

```
            )
         )
         //PREDICADO(i)#2#PREDV(i)
         (PREDICADO #é a língua oficial do Brasil#
            //PREDV(i)#14#SVTI(i) + OI
            (PREDV #é a língua oficial do Brasil#
               //SVTI(i)#1#[verbo(aux) + [verbo(aux)]] + verbo(vti(i))
               (SVTI #é#
                  (nucleo #é# verbo)
               )
               //OI(i)#4#OI_SIMPLES(i)
               (OI #a língua oficial do Brasil#
                  //OI_SIMPLES(i)#1#[poi] + SN(i)
                  (OI_SIMPLES #a língua oficial do Brasil#
                     //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                     (SN #a língua oficial do Brasil#
                        //AADNE(i)#2#SDET(i)
                        (AADNE #a#
                           //SDET(i)#1#[<todo>] + nucleo(art(i))
                           (SDET #a#
                              (nucleo #a# artigo)
                           )
                        )
                        (nucleo #língua# subst)
                        //AADND(i)#2#AADND_SIMPLES(i)
                        (AADND #oficial do Brasil#
                           //AADND_SIMPLES(i)#1#SADJ(i)
                           (AADND_SIMPLES #oficial do Brasil#
                              //SADJ(i)#1#[AADVL] + nucleo(adj(i)) + [CN] + [SADJ(i)]
                              (SADJ #oficial do Brasil#
                                 (nucleo #oficial# adj)
                                 //CN#2#SP
                                 (CN #do Brasil#
                                    //SP#1#preposicao + SN
                                    (SP #do Brasil#
                                       (nucleo #de# preposicao)
                                  //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                                       (SN #do Brasil#
                                          //AADNE(i)#2#SDET(i)
                                          (AADNE #do#
                                             //SDET(i)#1#[<todo>] + nucleo(art(i))
                                             (SDET #do#
                                                (nucleo #o# artigo)
                                             )
                                          )
                                          (nucleo #brasil# nome proprio)
                                       )
                                    )
                                 )
                              )
                           )
                        )
                     )
                  )
               )
            )
         )
      )
   )
)
```

3) >>>>>>>>>>>>>>>>>>>>>>>><<<<<<<<<<<<<<<<<<<<<<<<<<<

```
//FRASE#1#PERIODO
(FRASE #O português é a língua oficial do Brasil#
   //PERIODO#1#PERIODO_INDEPENDENTE
   (PERIODO #O português é a língua oficial do Brasil#
      //PERIODO_INDEPENDENTE#1#[AADVO] + SUJ(x) + [AADVO] + PREDICADO(x) + [AADVO]
      (PERIODO_INDEPENDENTE #O português é a língua oficial do Brasil#
         //SUJ(i)#2#SUJ_SIMPLES(i)
         (SUJ #O português#
            //SUJ_SIMPLES(i)#2#SN(i) + [APOSTO]
            (SUJ_SIMPLES #O português#
```

```
            //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
            (SN #O português#
               //AADNE(i)#2#SDET(i)
               (AADNE #O#
                  //SDET(i)#1#[<todo>] + nucleo(art(i))
                  (SDET #O#
                     (nucleo #o# artigo)
                  )
               )
               (nucleo #português# subst)
            )
         )
      )
      //PREDICADO(i)#1#PREDVN(i)
      (PREDICADO #é a língua oficial#
         //PREDVN(i)#3#SVTI(i) + OI(x) + POBJ(x)
         (PREDVN #é a língua oficial#
            //SVTI(i)#1#[verbo(aux) + [verbo(aux)]] + verbo(vti(i))
            (SVTI #é#
               (nucleo #é# verbo)
            )
            //OI(i)#4#OI_SIMPLES(i)
            (OI #a língua#
               //OI_SIMPLES(i)#1#[poi] + SN(i)
               (OI_SIMPLES #a língua#
                  //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                  (SN #a língua#
                     //AADNE(i)#2#SDET(i)
                     (AADNE #a#
                        //SDET(i)#1#[<todo>] + nucleo(art(i))
                        (SDET #a#
                           (nucleo #a# artigo)
                        )
                     )
                     (nucleo #língua# subst)
                  )
               )
            )
            //POBJ(i)#1#SADJ(i)
            (POBJ #oficial#
               //SADJ(i)#1#[AADVL] + nucleo(adj(i)) + [CN] + [SADJ(i)]
               (SADJ #oficial#
                  (nucleo #oficial# adj)
               )
            )
         )
      )
      //AADVO#4#AADVO_SIMPLES
      (AADVO #do Brasil#
         //AADVO_SIMPLES#2#SP
         (AADVO_SIMPLES #do Brasil#
            //SP#1#preposicao + SN
            (SP #do Brasil#
               (nucleo #de# preposicao)
               //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
               (SN #do Brasil#
                  //AADNE(i)#2#SDET(i)
                  (AADNE #do#
                     //SDET(i)#1#[<todo>] + nucleo(art(i))
                     (SDET #do#
                        (nucleo #o# artigo)
                     )
                  )
                  (nucleo #brasil# nome proprio)
               )
            )
         )
      )
   )
)
```

```
4) >>>>>>>>>>>>>>>>>>>>>>>>><<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<

//FRASE#1#PERIODO
(FRASE #O português é a língua oficial do Brasil#
   //PERIODO#1#PERIODO_INDEPENDENTE
   (PERIODO #O português é a língua oficial do Brasil#
      //PERIODO_INDEPENDENTE#1#[AADVO] + SUJ(x) + [AADVO] + PREDICADO(x) + [AADVO]
      (PERIODO_INDEPENDENTE #O português é a língua oficial do Brasil#
         //SUJ(i)#2#SUJ_SIMPLES(i)
         (SUJ #O português#
            //SUJ_SIMPLES(i)#2#SN(i) + [APOSTO]
            (SUJ_SIMPLES #O português#
               //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
               (SN #O português#
                  //AADNE(i)#2#SDET(i)
                  (AADNE #O#
                     //SDET(i)#1#[<todo>] + nucleo(art(i))
                     (SDET #O#
                        (nucleo #o# artigo)
                     )
                  )
                  (nucleo #português# subst)
               )
            )
         )
         //PREDICADO(i)#2#PREDV(i)
         (PREDICADO #é a língua oficial#
            //PREDV(i)#14#SVTI(i) + OI
            (PREDV #é a língua oficial#
               //SVTI(i)#1#[verbo(aux) + [verbo(aux)]] + verbo(vti(i))
               (SVTI #é#
                  (nucleo #é# verbo)
               )
               //OI(i)#4#OI_SIMPLES(i)
               (OI #a língua oficial#
                  //OI_SIMPLES(i)#1#[poi] + SN(i)
                  (OI_SIMPLES #a língua oficial#
                     //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                     (SN #a língua oficial#
                        //AADNE(i)#2#SDET(i)
                        (AADNE #a#
                           //SDET(i)#1#[<todo>] + nucleo(art(i))
                           (SDET #a#
                              (nucleo #a# artigo)
                           )
                        )
                        (nucleo #língua# subst)
                        //AADND(i)#2#AADND_SIMPLES(i)
                        (AADND #oficial#
                           //AADND_SIMPLES(i)#1#SADJ(i)
                           (AADND_SIMPLES #oficial#
                              //SADJ(i)#1#[AADVL] + nucleo(adj(i)) + [CN] + [SADJ(i)]
                              (SADJ #oficial#
                                 (nucleo #oficial# adj)
                              )
                           )
                        )
                     )
                  )
               )
            )
         )
         //AADVO#4#AADVO_SIMPLES
         (AADVO #do Brasil#
            //AADVO_SIMPLES#2#SP
            (AADVO_SIMPLES #do Brasil#
               //SP#1#preposicao + SN
               (SP #do Brasil#
                  (nucleo #de# preposicao)
                  //SN(i)#2#[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
                  (SN #do Brasil#
                     //AADNE(i)#2#SDET(i)
                     (AADNE #do#
                        //SDET(i)#1#[<todo>] + nucleo(art(i))
```

```
                        (SDET #do#
                           (nucleo #o# artigo)
                        )
                     )
                     (nucleo #brasil# nome proprio)
                  )
               )
            )
         )
      )
   )
)
Total de casamentos: 4
```

Menu pop-up - Trata-se de um menu acessado através do clique com o botão direito do mouse sobre a janela de visualização. Este menu permite a troca do modo de apresentação da janela de visualização.

Barra de status - Trata-se da barra cinza logo abaixo da janela de visualização. É utilizada para mostrar explicações mais detalhadas sobre alguns termos utilizados na interface e o status do teclado.

Guia de Ajuda - Trata-se de um manual eletrônico com informações dos principais tópicos referente ao CURUPIRA implementado em Microsoft HTML Help (Figura 4). Este manual possui todas as características de um hipertexto, permitindo a navegação, passando de um ponto a outro da mesma página ou de página diferente, usando os *links* de hipertexto.

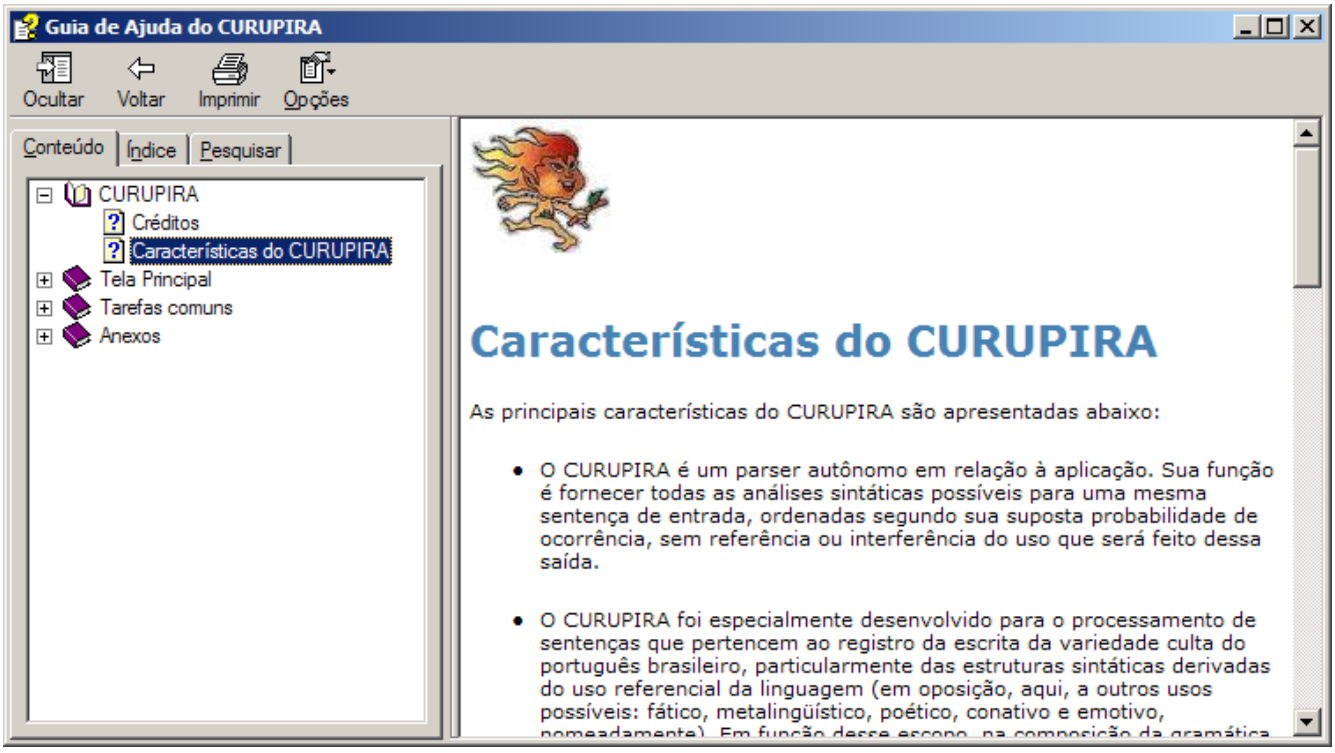


Figura 4: Guia de Ajuda do CURUPIRA

3.2. Procedimentos

O CURUPIRA processa a sentença em sentido descendente, recursivo, da esquerda para a direita, a partir da informação disponibilizada pelo dicionário e da prioridade de aplicação das regras estipulada pela gramática.

Para tanto, a sentença é processada seguindo os seguintes passos:

- *Tokenização* da sentença - a frase é quebrada em itens lexicais, e cada item lexical é classificado em um dos seguintes *tokens*:

```
T_FIM_STR: representa o fim da frase
T_ECOMERCIAL: token de & "E comercial"
T_PALAVRA: token de uma palavra
T_NUMERO: token de um numero
T_DOISPONTOS: token de dois pontos
T_PONTOEVIRGULA: token de ponto e virgula
T_PONTOFINAL: token de ponto final
T_VIRGULA: token de virgula
T_EXCLAMACAO: token de exclamação
T_INTERROGACAO: token de interrogação
T_TRESPONTOS: token de três pontos
T_SIMBOLOS: token de quaisquer outros símbolos
T_ABREPARENTESIS: token de abre parênteses
T_FECHAPARENTESIS: token de fecha parênteses
T_ABRECOLCHETE: token de abre colchetes
T_FECHACOLCHETE: token de fecha colchetes
T_ABRECHAVES: token de abre chaves
T_FECHACHAVES: token de fecha chaves
T_ASPAS: token de aspas
T_APOSTROFO: token de apostrofo
T_HIFEN: token de hífen
T_ENUMERACAO: token de enumeração
T_ABREVIATURA: token de abreviatura
T_REFERENCIA: token para "numero(numero)"
T_NUMEROREAL: token de um numero real
T_ACENTUACAO: token de acentuação (^´`¨~)
T_CONSOANTE: token de consoante
T_LETRANUMERO: token com mistura de letra e numero
T_NUMROMANO: token de numero romano
T_NUMFRACIONARIO: token de numero fracionário
T_DATA: token de data dia/mês/ano
T_NOMEARQUIVO: token para nomes de arquivos "nome.extensão"
T_INTERNET: token para endereço de internet
T_LIXO: token relacionado a um caractere estranho
```

- Recuperação das possíveis categorias e atributos dos itens lexicais do tipo palavra no dicionário compactado e verificação das mesóclises, ênclises e palavras compostas.

- Desambiguação lexical dos itens lexicais do tipo palavra que contenham mais de uma categoria distinta. Para cada item lexical é aplicada regras de desambiguação que utilizam, além das categorias do próprio item, informações dos itens lexicais vizinhos à esquerda e à direita;

- Analise Sintática dos itens lexicais – as regras do *parser* são disparadas enquanto houver sucesso nos casamentos das regras e itens lexicais. Zero ou mais casamentos entre as regras e itens lexicais são possíveis em função do mecanismo de *backtracking*;

- Os resultados são formatados e armazenados em um arquivo de saída percorrendo-se as árvores sintáticas formadas na memória do programa;

- Finalmente, o arquivo de saída é carregado pela interface para ser visualizado.

4. Trabalho Futuro

Dado o seu caráter incipiente, muito resta ainda a ser feito em relação ao CURUPIRA. Algumas dessas mudanças estão já em processo de desenvolvimento, como a que permite expandir o vocabulário terminal e não-terminal, evitando, assim, que o usuário precise acessar o *help* da ferramenta para poder interpretar a notação utilizada para registrar os resultados emitidos pela ferramenta. A possibilidade de edição on-line do conjunto de regras ou a customização do dicionário são outros atributos desejáveis ainda em processo de análise. De qualquer forma, alterações mais expressivas devem ser empreendidas assim que estiverem concluídos os testes a que a ferramenta vem sendo submetida.

5. Referências bibliográficas

[JAK63] Jakobson, R. Lingüística e poética. In Lingüística e Comunicação. São Paulo: Cultrix, 1995 (original publicado em inglês em 1960).

[KOW93] Kowaltowski, T.; Lucchesi, C.L. *Applications of Finite Automata Representing Large Vocabularies*. Software-Pratice and Experience, 23(1), 15-20, 1993.

[KOW95a] Kowaltowski, T.; Lucchesi, C.L.; Stolfi, J. *Minimization of Binary Automata*. Journal of the Brazilian Computing Society, 3(1), 36-42, 1995.

[KOW95b] Kowaltowski, T.; Lucchesi, C.L.; Stolfi, J. *Application of Finite Automata in Debugging Natural Language Vocabularies*. Journal of the Brazilian Computing Society, 3(1), 5-11, 1995.

[KOW98] Kowaltowski, T.; Lucchesi, C.L.; Stolfi, J. *Finite Automata and Efficient Lexicon Implementation.* Relatório Técnico IC-92-2, DCC/UNICAMP, 1998.

[NUN96] Nunes, M.G.V. et alli. *A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos*. Relatórios Técnicos do ICMC-USP, 42. Setembro 1996, 36p.

ANEXO I

VOCABULÁRIO NÃO-TERMINAL (em ordem alfabética)

LETRAS MAIÚSCULAS = VOCABULÁRIO NÃO-TERMINAL
letras minúsculas = vocabulário terminal ou quasi-terminal (categorias morfológicas)

AADND = adjunto adnominal à direita
AADND_COMPOSTO = adjunto adnominal à direita composto
AADND_SIMPLES = adjunto adnominal à direita simples
AADNE = adjunto adnominal à esquerda
AADVL = adjunto adverbial local
AADVL_COMPOSTO = adjunto adverbial local composto
AADVL_SIMPLES = adjunto adverbial local simples
AADVO = adjunto adverbial oracional
AADVO_COMPOSTO = adjunto adverbial oracional composto
AADVO_SIMPLES = adjunto adverbial oracional simples
AP = agente da passiva
APOSTO = aposto
CN = complemento nominal
coordenador = conjunto de conjunções coordenativas e outros elementos coordenadores (vírgula, etc.)
delimitador = delimitador de sentença
FRASE = frase
integrante = conjunção subordinativa integrante e outros elementos integrantes
nucleo(adj) = núcleo adjetivo
nucleo(adv) = núcleo advérbio
nucleo(artigo) = núcleo artigo
nucleo(subst) = núcleo substantivo
OD = objeto direto
OD_COMPOSTO = objeto direto composto
OD_SIMPLES = objeto direto simples
ODA = objeto direto anteposto
OI = objeto indireto
OI_COMPOSTO = objeto indireto composto
OI_SIMPLES = objeto indireto simples
OIA = objeto indireto anteposto
ORG = oração reduzida de gerúndio
ORI = oração reduzida de infinitivo
ORP = oração reduzida de particípio
OSADJ = oração subordinada adjetiva
OSAV = oração subordinada adverbial
OSSAP = oração subordinada substantiva agente da passiva
OSSAPO = oração subordinada substantiva apositiva
OSSCN = oração subordinada substantiva completiva nominal
OSSOD = oração subordinada substantiva objetiva direta
OSSOI = oração subordinada substantiva objetiva indireta
OSSPSUJ = oração subordinada substantiva predicativa do sujeito
OSSS = oração subordinada substantiva subjetiva
p = preposição
paadv = preposição que introduz adjunto adverbial

pap = preposição que introduz agente da passiva
pcn = preposição que introduz complemento nominal
PERÍODO = período
PERÍODO_COMPOSTO = período composto
PERÍODO_COORDENADO = período composto por coordenação
PERÍODO_INDEPENDENTE = período não-coordenado
PERÍODO_SIMPLES = período simples
POBJ = predicativo do objeto
poi = preposição que introduz objeto indireto
PREDICADO = predicado
PREDN = predicado nominal
PREDV = predicado verbal
PREDVN = predicado verbonominal
pron(subst) = pronome substantivo
pronome relativo = pronome relativo
PSUJ = predicativo do sujeito
PSUJ_COMPOSTO = predicativo do sujeito composto
PSUJ_SIMPLES = predicativo do sujeito simples
SADJ = sintagma adjetivo
SADV = sintagma adverbial
SDET = sintagma determinante
SN = sintagma nominal
SP = sintagma preposicional
SREL = sentença relativa
subordinante = conjunção subordinativa e outros elementos subordinantes
SUJ = sujeito
SUJ_COMPOSTO = sujeito composto
SUJ_SIMPLES = sujeito simples
SVI = sintagma verbal intransitivo
SVL = sintagma verbal de ligação
SVTD = sintagma verbal transitivo direto
SVTDI = sintagma verbal transitivo direto e indireto (bitransitivo)
SVTI = sintagma verbal transitivo indireto
verbo = verbo
verbo(aux) = verbo auxiliar
vi = verbo intransitivo
vtd = verbo transitivo direto
vtdi = verbo transitivo direto e indireto (bitransitivo)
vti = verbo transitivo indireto

ANEXO II

VOCABULÁRIO TERMINAL

Descrição da estrutura que armazena a categoria completa das palavras do dicionário.

As siglas utilizadas neste documento são:

M: Masculino
F: Feminino
2G: Dois Gêneros (Masculino e Feminino ao mesmo tempo)
S: Singular
P: Plural
2N: Dois Números (Singular e Plural ao mesmo tempo)
INV: Invariável
/: Indica um operador de Exclusividade (OR)
-: Nulo ou não consta classificação
(a,b,c,d): Indica um operador AND, ou seja, o verbete pode ser a,b,c e d ao mesmo tempo.

Cada verbete no dicionário terá 2 atributos:
1) Uma lista de possíveis categorias gramaticais, ORDENADAS POR FREQÜÊNCIA DE USO, e
2) Seus respectivos complementos (os complementos são conjuntos de outros campos, dependente da categoria).
Abaixo, apresentamos os campos com o tipo dos valores que podem receber (os campos do complemento aparecem seguindo cada categoria):

As categorias (classes) básicas do verbete são:

1- Substantivo

Gênero: M/F/2G/INV
Número: S/P/2N/INV
Grau: Aumentativo/Diminutivo/Nulo
Regência do Substantivo: (p_1, ..., p_n) uma lista (podendo ser nula) de preposições.
Regra Derivação Gênero: Identificador da regra (inteiro, por exemplo).
Regra Derivação Número: Identificador da regra.
Forma canônica: Uma palavra

Exemplos:
menino: M,S,-,-,1,2,menino (onde 1 é a regra em que o feminino é derivado trocando "o" final por "a". O identificador 2 acrescenta o "s" para formar o plural).
meninos: M,P,-,-,1,2,menino
meninão: M,S,Aumentativo,-,-,-,menino
lápis: M,2N,-,-,-,-,lápis
ajuda: F,S,-,(a),-,-,ajuda

2- Adjetivo

Gênero: M/F/2G/INV
Número: S/P/2N/INV
Grau: Aumentativo/Diminutivo/Superlativo/Nulo
Regência do Adjetivo: (p_1, ..., p_n) uma lista (podendo ser nula) de preposições.
Regra Derivação Gênero: Identificador da regra (inteiro, por exemplo).

Regra Derivação Número: Identificador da regra.
Forma canônica: Uma palavra
Exemplos:
bonito: M,S,-,-,1,2,bonito
bonitas: F,P,-,-,1,2,bonito
aprazível: 2G,S,-,-,-,3,aprazível (onde o identificador 3 troca "l" por "is" para formar o plural)
simples: 2G,2N,-,-,-,-,simples
igual: 2G,S,-,(a),-,3,igual
amabilíssimo: M,S,Superlativo,-,1,2,amável

3- Artigo
Gênero: M/F
Número: P/S
Tipo: Definido/Indefinido
Regra Derivação Gênero: Identificador da regra (inteiro, por exemplo).
Regra Derivação Número: Identificador da regra.
Forma canônica: Uma palavra

Exemplos:
o: M,S,Definido,1,2,o
umas: F,P,Indefinido,4,2,um (onde o identificador 4 acrescenta "a" para formar feminino)

4- Preposição
Contração: um par da forma (preposição, palavra), ou nula
Forma canônica: Uma palavra

Exemplos:
ante: -,ante
ao: (a, o), ao
do: (de, o), do
daqui: (de, aqui), daqui

5- Conjunção
Tipo: (Coordenativa,Subordinativa)
Complemento Coordenativa: (Aditiva, Adversativa,Alternativa, Conclusiva, Explicativa)
Complemento Subordinativa: (Integrante, Causal, Comparativa, Concessiva, Condicional, Consecutiva, Final, Temporal, Proporcional, Conformativa)
Forma canônica: Uma palavra
Exemplos:
nem: (Coordenativa), (Aditiva),-,nem
que: (Coordenativa, Subordinativa), (Aditiva, Alternativa), (Integrante, Causal, Comparativa, Final, Concessiva), que

6- Numeral
Gênero: M/F/2G/INV
Número: S/P/2N/INV
Tipo: Cardinal/Ordinal/Multiplicativo/Fracionário/Coletivo
Regra Derivação Gênero: Identificador da regra (inteiro, por exemplo).

Regra Derivação Número: Identificador da regra.
Forma canônica: Uma palavra

Exemplos:
segundo: M,S,Ordinal,1,2,segundo
duplo: M,S,Multiplicativo,1,2,duplo

7- Pronome

Gênero: M/F/2G/INV
Número: S/P/2N/INV
Tipo: (Pessoal Reto, Pessoal Oblíquo Átono, Pessoal Oblíquo Tônico, Possessivo, Demonstrativo, Indefinido, Interrogativo, Relativo, Reflexivo, Tratamento)
Regra Derivação Gênero: Identificador da regra (inteiro, por exemplo).
Regra Derivação Número: Identificador da regra.
Contração: um par da forma (preposição, palavra), ou nula
Forma canônica: Uma palavra

Exemplos:
Senhora: F,S,(Tratamento),5,2,-,senhora (onde o identificador 5 troca "a" por vazio para gerar o masculino)
eu: 2G,S, (Pessoal Reto),-,-,-,eu
dele: M,S, (Pessoal Reto),1,2, (de, ele),dele

8- Nomes Próprios

Gênero: M/F/2G
Número: S/P/2N
Forma canônica: Uma palavra
Exemplos:
Darci: 2G,S,Darci
Atlântico: M,S,Atlântico

9- Verbo

Predicação do Verbo: (Intransitivo, Transitivo Direto, Transitivo Indireto, Bitransitivo, Ligação, Auxiliar, Pronominal)
Formas Nominais: Infinitivo Pessoal/Gerúndio/Particípio/Nula
(no caso de *Particípio*, são necessárias as informações de gênero - M/F/2G e número - S/P/2N)
(no caso de *Infinitivo Pessoal*, o verbo possui Pessoa - eu, tu, ele, nós, vós, eles)
Par(Tempo,Pessoa):
Tempo: (Presente, Pretérito Imperfeito, Pretérito Perfeito, Pretérito Mais-que-Perfeito, Futuro do Presente, Futuro do Pretérito, Presente do Subjuntivo, Pretérito Imperfeito do Subjuntivo, Futuro do Subjuntivo, Imperativo Afirmativo) ou nulo
Pessoa: (eu, tu, ele, nós, vós, eles) ou nula
Regência do Verbo: (p_1, ..., p_n) uma lista (podendo ser nula) de preposições.
Forma canônica: Uma palavra

Obs: A primeira e a terceira pessoa do Infinitivo Pessoal coincidem com o Infinitivo Impessoal. Assim, não colocamos a forma nominal Infinitivo Impessoal como complemento do verbo.

Exemplos:
meditar: (Intransitivo,Transitivo Indireto, Transitivo Direto),-,((Infinitivo Pessoal, eu); (Infinitivo Pessoal, ele), (Futuro do Subjuntivo, eu); (Futuro do Subjuntivo, ele)), (em, sobre), meditar
vendido: (Intransitivo,Transitivo Direto, Bitransitivo), (Particípio,M,S),-,vender

10- Advérbio

Tipo: Circunstância Lugar/ Circunstância Tempo/ Circunstância Modo/ Negação/ Dúvida/ Intensidade/ Afirmação/ Interrogativo de Tempo/ Interrogativo de Modo/ Interrogativo de Causa/ Interrogativo de Lugar
Grau: Aumentativo/Diminutivo/Superlativo/Nulo
Forma canônica: Uma palavra
Exemplos:
abaixo: Circunstância Lugar,-,abaixo
onde: Interrogativo de Lugar,-,onde

11.- Prefixos

Forma canônica: Uma palavra
Exemplos:
super: super
pós: pós
sub: sub

12- Siglas

Forma canônica: Uma palavra
Exemplos:
ONU: ONU
PDT: PDT
OTAN: OTAN
USP: USP

13- Abreviaturas

Gênero: M/F/2G
Número: S/P/2N
Regra Derivação Gênero: Identificador da regra (inteiro, por exemplo).
Regra Derivação Número: Identificador da regra.
Forma canônica: Uma palavra

Exemplos:
S.Exa.: 2G,S,-,-,S.Exa.
S.Ema.: 2G,S,-,-,S.Ema.

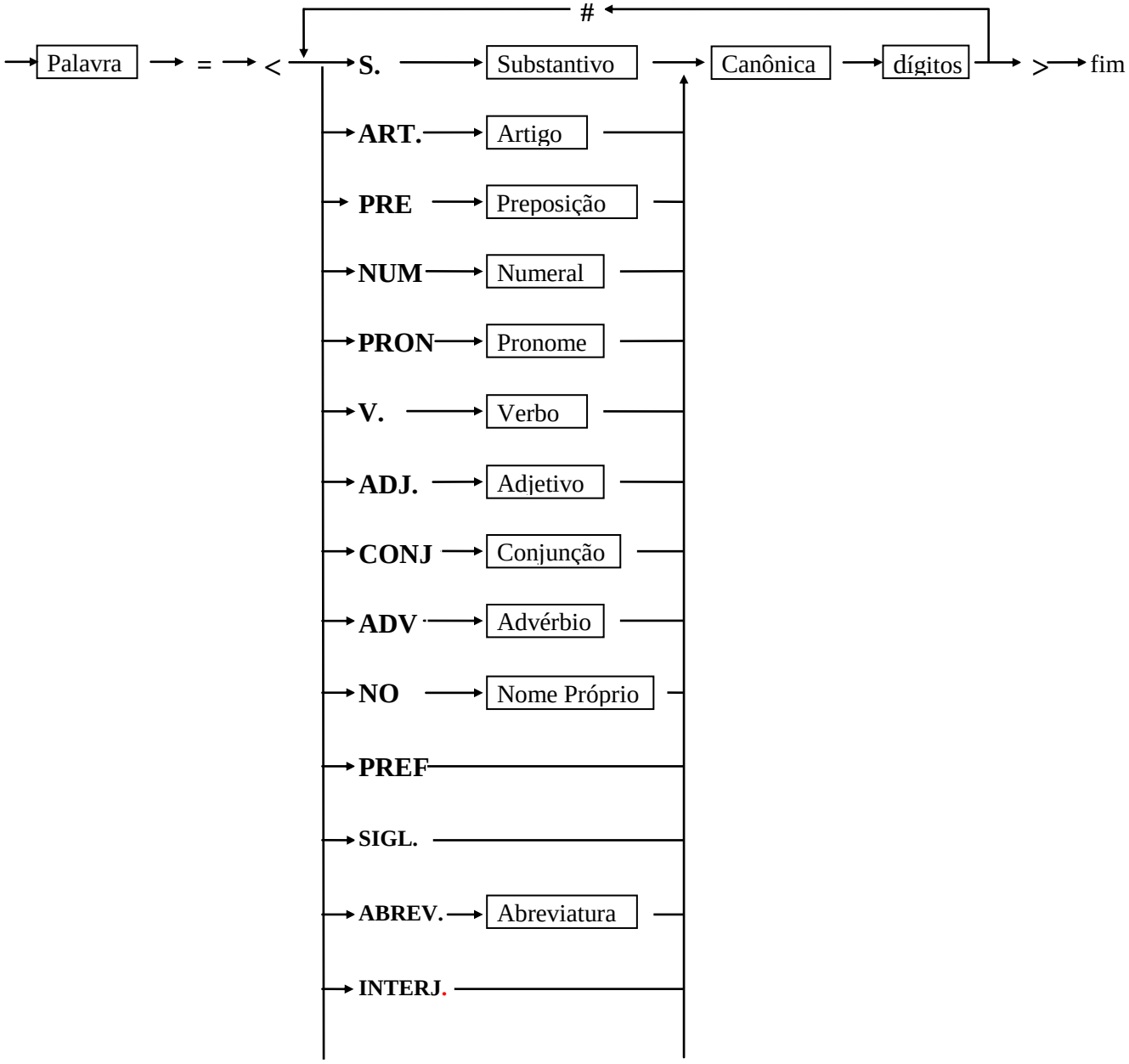
14- Interjeição

Forma canônica: Uma palavra
Exemplos:
Ah:Ah
Ih: Ih
Olá: Olá
Oi: Oi

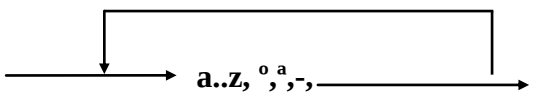
ANEXO IIB - ABREVIATURAS UTILIZADAS NA SINTAXE

A. = aumentativo
ABREV. = abreviatura
ADIT. = aditiva
ADJ. = adjetivo
ADVE. = adversativa
AFIR. = afirmação
ALTER. = alternativa
ART. = artigo
AUX. = auxiliar
BI. = bitransitivo
C. = contração
CAR. = cardinal
CAUS. = causal
CIR_LUG. = circunstância lugar
CIR_TEMP. = circunstância tempo
CIR_MOD. = circunstância modo
CO. = coletivo
COMP. = comparativa
CONC. = concessiva
CONCL. = conclusiva
COND. = condicional
CONFOR. = conformativa
CONS. = consecutiva
COORD. = coordenativa
D. = diminutivo
DE. = definido
DEM. = demonstrativo
DUV. = dúvida
ENC. = ênclise
EXPL. = explicativa
F. = feminino
FIN. = final
FRA. = fracionário
FUT_PRES. = futuro do presente
FUT_SUBJ. = futuro do subjuntivo
GERUN. = gerúndio
IMP_AFIRM. = imperativo afirmativo
INF_PESS. = infinitivo pessoal
INDE. = indefinido
INT. = intransitivo/intensidade
INTE. = integrante
INTER. = interrogativo
INTERJ. = interjeição
INT_CAUS. = interrogativo de causa
INT_LUG. = interrogativo de lugar
INT_MOD. = interrogativo de modo
INT_TEMP. = interrogativo de tempo
INV. = invariável
LIG. = ligação
M. = masculino
MESC. = mesóclise
MUL. = multiplicativo
N. = nulo
NEG. = negação
NOM. = nome próprio
NUM . = numeral
PARTIC. = particípio
PL. = plural
POSS. = possessivo
PREF. = prefixo
PREP. = preposição
PRES. = presente
PRES_SUBJ. = presente do subjuntivo
PRET_IMPERF. = pretérito imperfeito
PRET_IMPERF_SUBJ. = pretérito imperfeito do subjuntivo
PRET_M_Q_P. = pretérito mais que perfeito
PRET_PERF. = pretérito perfeito
PRON. = pronome
PRONOM.= pronominal
PROPOR. = proporcional
RET. = pessoal reto
REL. = relativo
REFL. = reflexivo
S. = substantivo
SI. = singular
SIGL. = sigla
SU. = superlativo
SUBORD. = subordinativa
TD. = transitivo direto
TEMP. = temporal
TI. = transitivo indireto
TRAT. = tratamento
V. = verbo
2G. = dois gênero
2N. = dois número

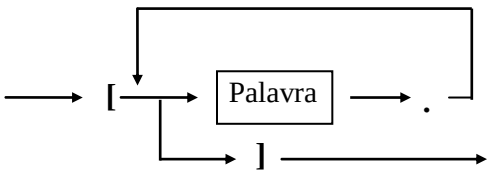
Início:



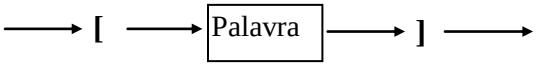
Palavra:



Lista de Palavras:

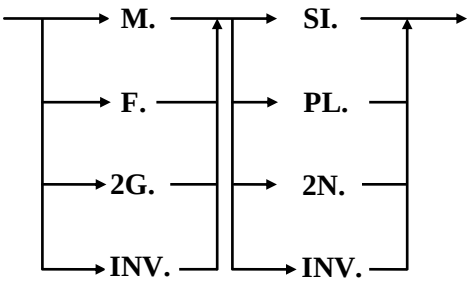


Canônica:

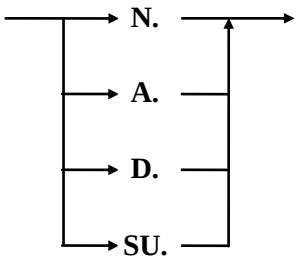


Regência:

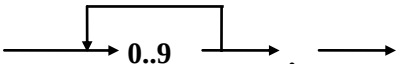
Gênero e Número:



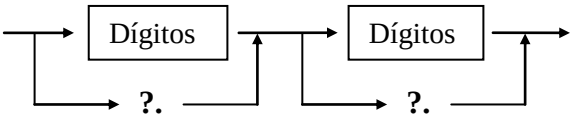
Grau:



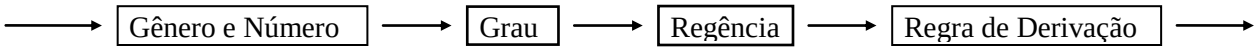
Dígitos:



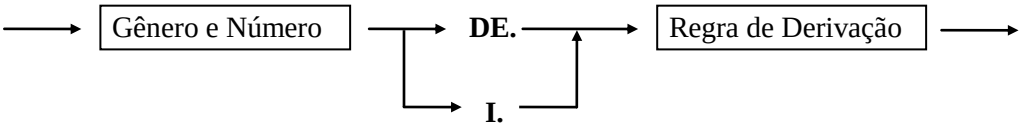
Regra de Derivação:



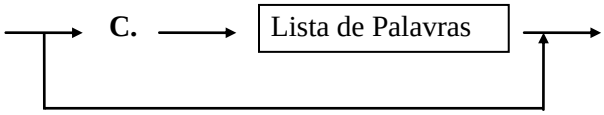
Substantivo ou Adjetivo:



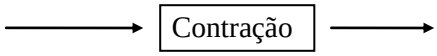
Artigo:



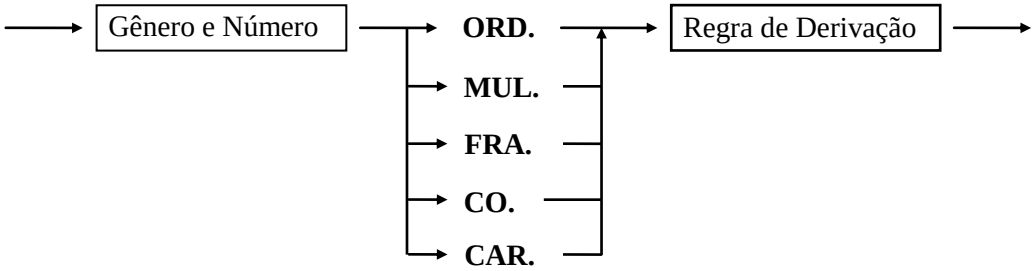
Contração:



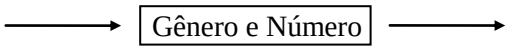
Preposição:



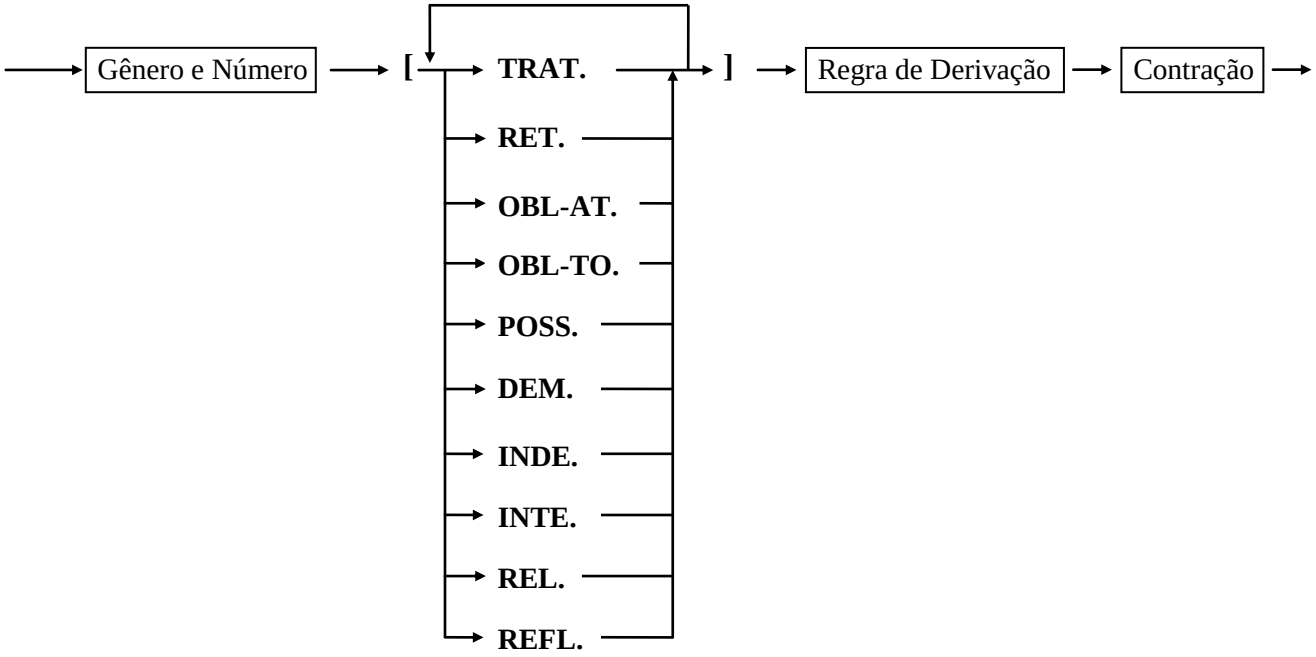
Numeral:



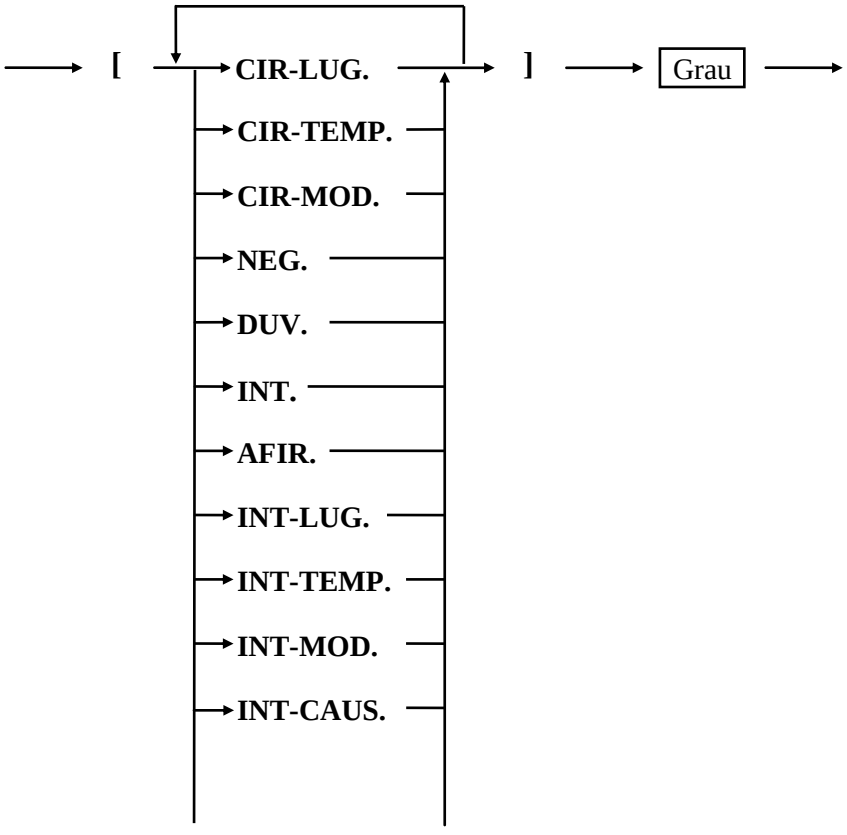
Nome Próprio:



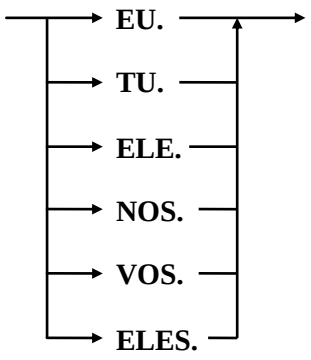
Pronome:



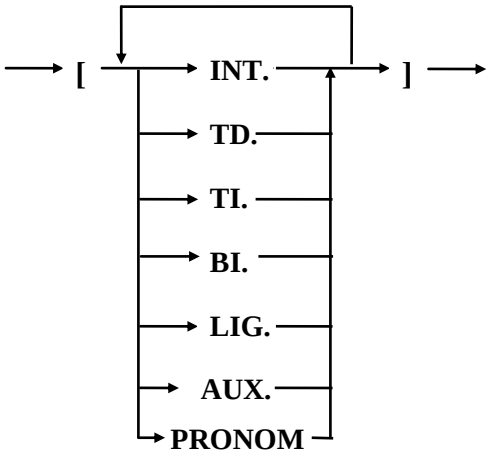
Advérbio:



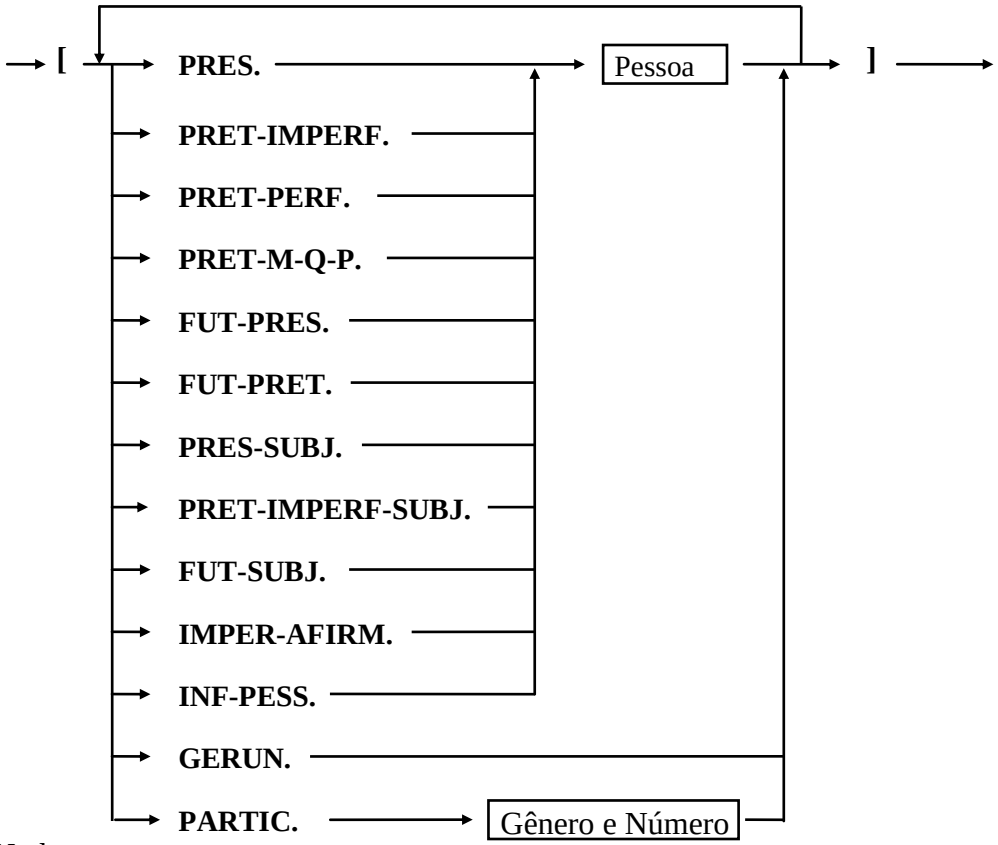
Pessoa:



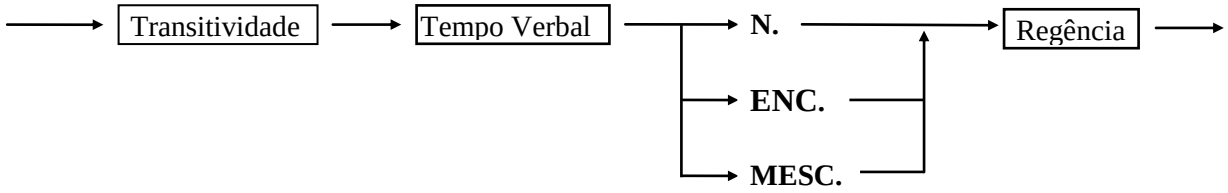
Transitividade:



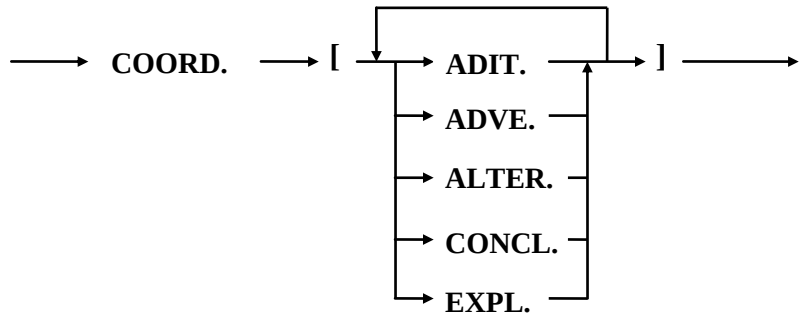
Tempo Verbal:



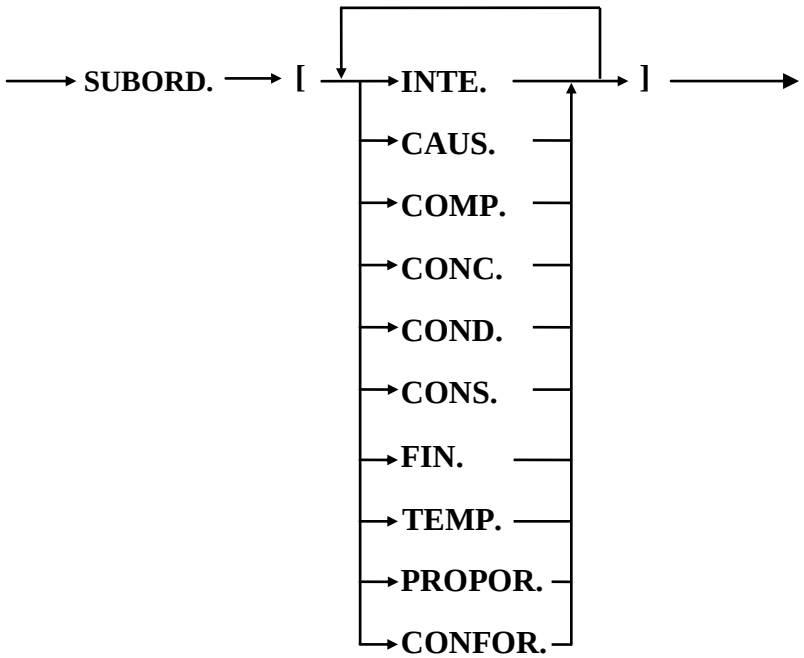
Verbo:



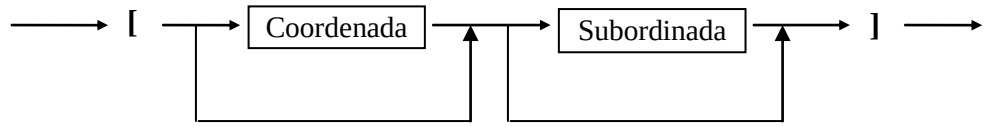
Coordenada:



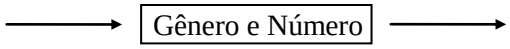
Subordinada:



Conjunção:



Abreviatura:



ANEXO III

REGRAS DE REESCRITA CATEGORIAL

CHAVE DE INTERPRETAÇÃO

LETRAS MAIÚSCULAS = VOCABULÁRIO NÃO-TERMINAL
letras minúsculas = vocabulário terminal ou quasi-terminal (categorias morfológicas)
+ = justaposição, com separação por espaço em branco (apenas).
[X] = opcionalidade (x é opcional)
X(Y) = Y deve ser atributo de X
(ii) = indexado (os termos portadores de (i) devem concordar em número, pessoa e gênero)
{X,Y} = exclusividade: ou X, ou Y

1. Definição da frase

CATEGORIA	P	EXPANSÃO
FRASE	1	PERIODO
FRASE	2	delimitador + [AADVO] + SUJ + [AADVO] + delimitador

2. Definição do período

CATEGORIA	P	EXPANSÃO
PERIODO	1	PERIODO_COORDENADO
PERIODO	2	PERIODO_INDEPENDENTE
PERIODO_COORDENADO	1	[coordenador] + PERIODO_INDEPENDENTE + coordenador + PERIODO_COORDENADO
PERIODO_COORDENADO	2	[coordenador] + PERIODO_INDEPENDENTE + coordenador + PERIODO_INDEPENDENTE
PERIODO_INDEPENDENTE	1	[AADVO] + SUJ(i) + [AADVO] + PREDICADO(i) + [AADVO]
PERIODO_INDEPENDENTE	2	[AADVO] + PREDICADO(i) + [AADVO] + SUJ(i) + [AADVO]
PERIODO_INDEPENDENTE	3	[AADVO] + PREDICADO + [AADVO]

3. Definição dos termos essenciais da oração

3.1 Sujeito

CATEGORIA	P	EXPANSÃO
SUJ	1	OSSS
SUJ	2	SUJ_COMPOSTO
SUJ	3	SUJ_SIMPLES
SUJ_COMPOSTO	1	[coordenador] + SUJ_SIMPLES + coordenador + SUJ_COMPOSTO
SUJ_COMPOSTO	2	[coordenador] + SUJ_SIMPLES + coordenador + SUJ_SIMPLES
SUJ_SIMPLES	1	{eu, tu, ele, ela, nós, vós, eles, elas}
SUJ_SIMPLES	2	SN

3.2 Predicado

CATEGORIA	P	EXPANSÃO
PREDICADO	1	PREDVN
PREDICADO	2	PREDV
PREDICADO	3	PREDN

3.2.1 Predicado Nominal

CATEGORIA	P	EXPANSÃO
PREDN	1	SVL + PSUJ(i=SUJ)

3.2.2 Predicado Verbal

CATEGORIA	P	EXPANSÃO
PREDV	1	verbo(aux) + [verbo(aux)] + ODA + SVTDI + OI
PREDV	2	verbo(aux) + [verbo(aux)] + OIA + SVTDI + OD
PREDV	3	verbo(aux) + [verbo(aux)] + ODA + SVTD
PREDV	4	verbo(aux) + [verbo(aux)] + OIA + SVTI
PREDV	5	ODA + SVTDI + OI
PREDV	6	OIA + SVTDI + OD
PREDV	7	ODA + SVTD
PREDV	8	OIA + SVTI
PREDV	9	SVTDI + OD + OI
PREDV	10	SVTDI + OI + AP
PREDV	11	SVTDI + OI + OD
PREDV	12	SVTD + AP
PREDV	13	SVTD + OD
PREDV	14	SVTI + OI
PREDV	15	SVI

3.2.3 Predicado Verbo-Nominal

CATEGORIA	P	EXPANSÃO
PREDVN	1	ODA(i) + SVTD + POBJ(i)
PREDVN	2	ODA + SVTD + PSUJ(i=SUJ)
PREDVN	3	SVTI + OI(i) + POBJ(i)
PREDVN	4	SVTD + OD(i) + POBJ(i)
PREDVN	5	SVTD + POBJ(i) + OD(i)
PREDVN	6	SVTD + OD + PSUJ(i=SUJ)
PREDVN	7	SVI + PSUJ(i=SUJ)

3.3 Predicativo

3.3.1 Predicativo do Objeto

CATEGORIA	P	EXPANSÃO
POBJ	1	SADJ
POBJ	2	SN

3.3.2 Predicativo do Sujeito

CATEGORIA	P	EXPANSÃO
PSUJ	1	OSSPSUJ
PSUJ	2	PSUJ_COMPOSTO
PSUJ	3	PSUJ_SIMPLES
PSUJ_COMPOSTO	1	[coordenador] + PSUJ_SIMPLES + coordenador + PSUJ_COMPOSTO
PSUJ_COMPOSTO	2	[coordenador] + PSUJ_SIMPLES + coordenador + PSUJ_SIMPLES
PSUJ_SIMPLES	1	SADJ
PSUJ_SIMPLES	2	SN

4. Definição dos termos integrantes da oração

4.1 Complemento Nominal

CATEGORIA	P	EXPANSÃO
CN	1	OSSCN
CN	2	SP

4.2 Complemento Verbal

4.2.1 Objeto Direto

CATEGORIA	P	EXPANSÃO
OD	1	hifen + {me, te, se, o, a, nos, vos, os, as}
OD	2	OSSOD
OD	3	OD_COMPOSTO
OD	4	OD_SIMPLES
OD_COMPOSTO	1	[coordenador] + OD_SIMPLES + coordenador + OD_COMPOSTO
OD_COMPOSTO	2	[coordenador] + OD_SIMPLES + coordenador + OD_SIMPLES
OD_SIMPLES	1	SN
ODA	1	{me, te, se, o, a, nos, vos, os, as}

4.2.2 Objeto Indireto

CATEGORIA	P	EXPANSÃO
OI	1	hifen + {me, te, se, lhe, nos, vos, lhes}
OI	2	poi + {mim, ti, si, ele, ela, nós, vós, eles, elas}
OI	3	OSSOI
OI	4	OI_COMPOSTO
OI	5	OI_SIMPLES
OI_COMPOSTO	1	[coordenador] + OI_SIMPLES + coordenador + OI_COMPOSTO
OI_COMPOSTO	2	[coordenador] + OI_SIMPLES + coordenador + OI_SIMPLES
OI_SIMPLES	1	[poi] + SN
OIA	1	{me, te, se, lhe, nos, vos, lhes}

4.3 Agente da Pasiva

CATEGORIA	P	EXPANSÃO
AP	1	OSSAP
AP	2	pap + SN

5. Definição dos termos acessórios da oração

5.1. Adjunto Adnominal

CATEGORIA	P	EXPANSÃO
AADND	1	OSADJ
AADND	2	AADND_COMPOSTO
AADND	3	AADND_SIMPLES
AADND_COMPOSTO	1	[coordenador] + AADND_SIMPLES + coordenador + AADND_COMPOSTO
AADND_COMPOSTO	2	[coordenador] + AADND_SIMPLES + coordenador + AADND_SIMPLES
AADND_SIMPLES	1	SADJ
AADND_SIMPLES	2	SP
AADNE	1	[SDET] + nucleo(adj)
AADNE	2	SDET

5.2 Adjunto Adverbial Oracional

CATEGORIA	P	EXPANSÃO
AADVO	1	OSADV
AADVO	2	{comigo, contigo, consigo, conosco, convosco}
AADVO	3	paadv + {mim, ti, si, ele, ela, nós, vós, eles, elas}
AADVO	4	AADVO_COMPOSTO
AADVO	5	AADVO_SIMPLES
AADVO_COMPOSTO	1	[coordenador] + AADVO_SIMPLES + coordenador + AADVO_COMPOSTO
AADVO_COMPOSTO	2	[coordenador] + AADVO_SIMPLES + coordenador + AADVO_SIMPLES
AADVO_SIMPLES	1	SADV
AADVO_SIMPLES	2	SP

5.3 Adjunto Adverbial Local

CATEGORIA	P	EXPANSÃO
AADVL	2	{comigo, contigo, consigo, conosco, convosco}
AADVL	3	paadv + {mim, ti, si, ele, ela, nós, vós, eles, elas}
AADVL	4	AADVL_COMPOSTO
AADVL	5	AADVL_SIMPLES
AADVL_COMPOSTO	1	[coordenador] + AADVL_SIMPLES + coordenador + AADVL_COMPOSTO
AADVL_COMPOSTO	2	[coordenador] + AADVL_SIMPLES + coordenador + AADVL_SIMPLES
AADVL_SIMPLES	1	SADV

5.4 Aposto

CATEGORIA	P	EXPANSÃO
APOSTO	1	OSSAPO
APOSTO	2	SN

6. Definição dos sintagmas básicos

6.1 Sintagma Adjetivo

CATEGORIA	P	EXPANSÃO
SADJ	1	[AADVL] + nucleo(adj(i)) + [CN] + [SADJ(i)]

6.2 Sintagma Adverbial

CATEGORIA	P	EXPANSÃO
SADV	1	[AADVL] + nucleo(adv) + [CN] + [AADVL]

6.3 Sintagma Determinante

CATEGORIA	P	EXPANSÃO
SDET	1	[<todo>] + nucleo(artigo)
SDET	2	{cerca de, perto de, mais de}
SDET	3	pronome relativo

6.4 Sintagma Nominal

CATEGORIA	P	EXPANSÃO
SN	1	pron(subst)
SN	2	[AADVL] + [AADNE(i)] + nucleo(subst(i)) + [CN] + [AADND(i)]
SN	3	[AADVL] + AADNE(i) + nucleo(adj(i)) + [AADND(i)]

6.5 Sintagma Preposicional

CATEGORIA	P	EXPANSÃO
SP	1	p + SN
SP	2	p + SADJ
SP	3	p + SADV

6.5 Sintagma Verbal

CATEGORIA	P	EXPANSÃO
SVL	1	[AADVL] + [verbo(aux) + [verbo(aux)]] + vl
SVTD	1	[AADVL] + [verbo(aux) + [verbo(aux)]] + vtd
SVTDI	1	[AADVL] + [verbo(aux) + [verbo(aux)]] + vtdi
SVTI	1	[AADVL] + [verbo(aux) + [verbo(aux)]] + vti
SVI	1	[AADVL] + [verbo(aux) + [verbo(aux)]] + vi

7. Definição dos sintagmas oracionais

7.1 Oração Subordinada Adjetiva

CATEGORIA	P	EXPANSÃO
OSADJ	1	ORG
OSADJ	2	ORP
OSADJ	3	SREL

7.2 Oração Subordinada Adverbial

CATEGORIA	P	EXPANSÃO
OSADV	1	ORG
OSADV	2	ORP
OSADV	3	subordinante + PERIODO

7.3 Oração Subordinada Substantiva

CATEGORIA	P	EXPANSÃO
OSS	1	ORI
OSS	2	integrante + PERIODO
OSSAPO	1	PERIODO
OSSCN	1	[pcn] + OSS
OSSOD	1	OSS
OSSOI	1	[poi] + OSS
OSSPSUJ	1	OSS
OSSS	1	OSS
OSSAP	1	OSS

7.4 Orações Reduzidas

CATEGORIA	P	EXPANSÃO
ORI	1	PERIODO(verbo = infinitivo)
ORP	1	PERIODO(verbo = particípio)
ORG	1	PERIODO(verbo = gerúndio)

8. Definição das orações relativas

CATEGORIA	P	EXPANSÃO
SREL	1	PERIODO(sujeito = pronome relativo)
SREL	2	PERIODO(od = pronome relativo)
SREL	3	PERIODO(oi = pronome relativo)
SREL	4	PERIODO(aadv = pronome relativo)
SREL	5	PERIODO(aadne = pronome relativo)
SREL	6	PERIODO(cn = pronome relativo)

9. Definição do vocabulário terminal (letras minúsculas):

CATEGORIA	P	EXPANSÃO
coordenador	1	[virgula] + conjuncao coordenativa
coordenador	2	virgula
coordenador	3	ponto-e-virgula
delimitador	1	reticências
delimitador	2	ponto de interrogação
delimitador	3	ponto de exclamação
delimitador	4	dois pontos
delimitador	5	ponto final
delimitador	6	marcador de fim de parágrafo
delimitador	7	marcador de tabulação
delimitador	8	marcador de fim de linha
integrante	1	conjuncao integrante
nucleo(adj)	1	adjetivo
nucleo(adj)	2	numeral cardinal
nucleo(adj)	3	numeral ordinal
nucleo(adj)	4	pronome demonstrativo variável
nucleo(adj)	5	pronome indefinido variável
nucleo(adj)	6	pronome possessivo
nucleo(adj)	7	algarismo arábico
nucleo(adv)	1	advérbio
nucleo(adv)	1	advérbio
nucleo(subst)	1	substantivo
nucleo(subst)	1	nome próprio
nucleo(subst)	1	toda e qualquer palavra desconhecida
nucleo(subst)	1	sigla
nucleo(subst)	1	abreviatura
nucleo(subst)	1	numeral multiplicativo
nucleo(subst)	1	numeral fracionário
nucleo(subst)	1	numeral coletivo
p	1	preposicao
paadv	1	p
pap	1	{<de>, <por>}
pcn	1	{<de>, , com}
poi	1	{<de>, , com, para}
pron(subst)	1	pronome demonstrativo invariavel
pron(subst)	1	pronome indefinido invariavel
pron(subst)	1	pronome interrogativo
pron(subst)	1	pronome relativo
subordinante	1	conjuncao subordinativa
vi	1	verbo intransitivo
vl	1	verbo de ligação
vtd	1	verbo transitivo direto

vtdi	1	verbo transitivo direto e indireto
vti	1	verbo transitivo indireto
verbo(aux)	1	verbo auxiliar
verbo(aux)	2	{<começar> a, <terminar> de, <continuar> a, <deixar> de}