Cadernos de
ESTUDOS LINGUÍSTICOS – (58.2), Campinas, pp. 299-315 - mai./ago. 2016

CRIANDO "BANCOS DE ÁRVORES": O SISTEMA DE ANOTAÇÃO E O PROCESSAMENTO AUTOMÁTICO

PABLO FARIA[1]
CHARLOTTE GALVES[2]

RESUMO. Neste trabalho, assinalamos a estreita relação entre sistemas de anotação e análise sintática automática, apresentando um experimento para comparar análises automáticas baseadas nas versões atual e modificada do sistema de etiquetas morfológicas verbais utilizado no *Corpus Tycho Brahe*. A modificação resultou em um ganho aproximado de dois pontos percentuais na medida F1 de acurácia, conforme medida pelo aplicativo *evalb*. Este resultado indica que o sistema de anotação pode ser pensado de modo a ser mais conciso e informativo ao analisador sintático automático (doravante, analisador). Como conclusão, são sugeridos dois princípios norteadores para especificação do sistema de anotação e treinamento do analisador. Por fim, a discussão é contextualizada por uma visão geral e uma breve discussão do processo de construção de "bancos de árvores" (*treebanks*) e de sua importância na pesquisa linguística.
Palavras-chave: linguística de corpus, corpora anotados, processamento automático

ABSTRACT. In this paper, we highlight the tight relation between annotation systems and parsing by presenting an experiment for evaluation of alternative parses based on current and modified versions of the verbal tag system used in the *Tycho Brahe Corpus*. The modified version resulted in an improvement of two percentage points in the F1 measure of parsing accuracy, as evaluated by the *evalb* software. This result shows that the annotation system can be devised in order to be more concise and informative to the parser. As a conclusion, we suggest two guidelines for the specification of annotation systems and the training of the parser. Finally, the present discussion is contextualized by an outline and a brief discussion of the process of treebank building and of its importance for linguistic research.
Keywords: corpus linguistics, annotated corpora, automatic processing

[1] Professor na Universidade Estadual de Campinas. e-mail: pablofaria@iel.unicamp.br. O presente trabalho foi conduzido no âmbito do projeto de pesquisa FAPESP 13/18090-6.

[2] Professora na Universidade Estadual de Campinas. e-mail: galvesc@unicamp.br. Coordenadora do projeto temático FAPESP 12/06078-9, ao qual estão vinculados este estudo e o projeto de pesquisa mencionado acima.

1. INTRODUÇÃO

É crescente a utilização de corpora linguísticos (textos, transcrições de fala, de bate-papos de internet etc.) para estudos sobre a linguagem. Particularmente, no âmbito dos estudos diacrônicos sobre sintaxe das línguas, vem crescendo o número de "bancos de árvores" (adaptado do termo em inglês, *treebank*[3]), que são corpora de dados linguísticos transcritos, enriquecidos com anotação de informações sintáticas e/ou semânticas, na forma de representações arbóreas em que se indicam as relações entre elementos no interior de sentenças ou fragmentos de sentenças. Embora a expressão *banco de árvores* remeta fortemente, nos dias atuais, à iniciativa de Mitchell Marcus e colegas (Marcus et al, 1993) que resultou na criação do primeiro banco de árvores em larga escala (i.e., na casa dos milhões de palavras), o *Penn Treebank* (Taylor et al., 2003), esse termo tem um escopo mais amplo.

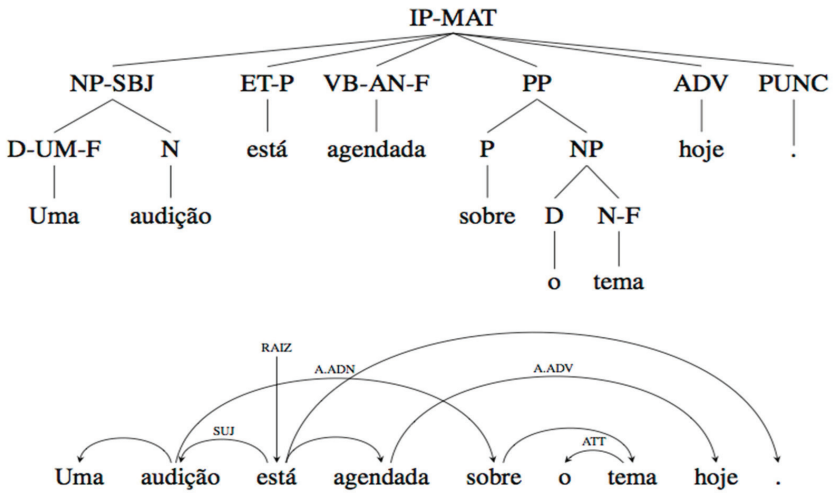

Figura 1. Exemplos de árvores: de estrutura sintagmática[4] e de dependência.

Há dois tipos principais de bancos de árvores sintáticos: os de estrutura sintagmática (*phrase structure*), como o *Corpus Tycho Brahe* (doravante, CTB) (Galves & Faria, 2010), e os de dependência (*dependency structure*), como o *Chinese Dependency Treebank 1.0* (Che et al, 2012), entre outros. Para além destes, há ainda bancos de árvores baseados em Gramática Categorial Combinatória, HPSG, Gramática de Caso, LFG e outras que integram estrutura sintagmática e dependência (ver Xiao, 2008, para um levantamento de vários *corpora*).

[3] O termo *treebank* - correspondente a "banco de árvore", em analogia a "banco de sangue" - foi cunhado nos anos 1980, na Universidade de Lancaster, por pesquisadores que incluíam Geoffrey Neil Leech e Geoffrey Sampson, este último tendo sido o primeiro a construir um banco de árvores (Leech, 2009).

[4] Exemplo criado com base no sistema de anotação do *Corpus Tycho Brahe* (Galves & Faria, 2010).

A Figura 1 exemplifica os dois tipos principais de bancos de árvores. Note-se que há semelhanças entre as duas representações, quanto às interdependências dos elementos. Porém, não são representações completamente isomórficas, o que se vê, por exemplo, na forma como o advérbio "hoje" aparece nas estruturas: na primeira, não é indicado qual dos verbos ele modifica, visto que ele é apenas um nó "irmão" dos nós relativos aos verbos auxiliar e principal.

Outra diferença importante é a de que, enquanto as árvores de dependência são explícitas quanto a quais itens lexicais são os núcleos a partir dos quais dependências são determinadas, nas árvores de estrutura sintagmática esta informação é interpretada com base nos rótulos dos nós sintáticos e nas etiquetas morfossintáticas dos itens lexicais, interpretação que depende do conhecimento prévio do sistema de anotação respectivo.

Em princípio, um banco de árvores pode aplicar qualquer formalismo gramatical cuja estrutura atribuída às sentenças seja arbórea. Além disso, bancos de árvores de um mesmo tipo podem variar bastante entre si, seja quanto a quais propriedades gramaticais são anotadas, seja em como as anotam. Por exemplo, na teoria da Gramática Gerativa, postulam-se elementos invisíveis nas sentenças, chamados de "categorias vazias", o que leva certos bancos de árvores a codificar tais informações. No CTB, por exemplo, são anotadas ocorrências de sujeito nulo, relações de longa distância entre elementos interrogativos e as posições em que são interpretados, entre outras propriedades. O grau de compromisso com teorias e análises linguísticas específicas também varia bastante entre os *corpora*.

Desde que o desenvolvimento de bancos de árvores em larga escala se intensificou, a partir dos anos 1990, seu impacto na pesquisa linguística e também na área de linguística computacional tem sido cada vez mais significativo, como mostramos nas próximas seções. Neste artigo, é abordada a relação entre o sistema de anotação utilizado no CTB e o processamento automático, em particular, no que diz respeito à análise sintática automática. Nosso objetivo central é clarificar um pouco mais este aspecto da construção de bancos de árvores e divisar algumas diretrizes que auxiliem na especificação ou na revisão de sistemas de anotação, de modo a impactar positivamente o desempenho da análise sintática automática.

O artigo está organizado da seguinte forma: na **seção 2**, é feita uma introdução ao processo de anotação sintática para criação de um banco de árvores. Na **seção 3**, apontamos a importância dos bancos de árvores para a pesquisa em linguística e, em particular, para a linguística histórica. A **seção 4** adentra o campo da linguística computacional, discutindo a relação entre anotação e análise sintática automática com base em experimentos com o analisador de Dan Bikel (Bikel, 2004). Os experimentos são apresentados e seus resultados discutidos à luz de estudos similares na literatura. A **seção 5** sintetiza e organiza as conclusões tiradas da discussão na seção anterior, de modo a sugerir diretrizes para especificação de sistemas de anotação. Por fim, a **seção 6** apresenta conclusões gerais e observações finais.

2. A CONSTRUÇÃO DE BANCOS DE ÁRVORES

A construção de bancos de árvores consiste em acrescentar informação aos dados linguísticos brutos, isto é, em que há apenas a transcrição das sentenças e fragmentos. A anotação sintática pode ser vista, portanto, como o "enriquecimento" de um *corpus* (Lavid, 2013) que, do ponto de vista da área de processamento de linguagem natural (PLN), consiste na transformação de texto puro em texto *marcado* (e, de certo modo, interpretado). A disponibilidade deste tipo de recurso é fundamental para várias aplicações na linguística computacional, tais como a extração de informação, sumarização de textos, tradução etc.

Esta construção, atualmente, é semiautomática e envolve, geralmente, programas de computador que implementam algoritmos de aprendizagem (de máquina). Tais aplicativos, que chamaremos simplificadamente de analisadores (como adaptação de *parsers*, que se referem a analisadores sintáticos automáticos), são treinados sobre uma quantidade razoável de dados pré-anotados para aprender o sistema de anotação e aplicá-lo a novos textos. Essa anotação automática é imperfeita, entretanto, o que implica revisão humana. Nesta seção, apresentamos uma visão geral do processo de anotação, segundo o que é delineado por Lavid (2013):

1. **Seleção de textos representativos.** O primeiro passo para a construção de um banco de árvores envolve a seleção dos textos representativos segundo alguma hipótese ou interesse de pesquisa.

2. **Especificação do sistema de anotação**. Uma vez definidos os textos que constituirão o *corpus*, passa-se à definição da teoria linguística que determinará a forma de interpretação dos dados e orientará a especificação do sistema de anotação. Nesta fase, começa-se também a produção do manual de anotação, em que se descreve o sistema, suas motivações e assunções, e que será utilizado para treinamento de anotadores/revisores humanos.

3. **Teste do sistema de anotação**. Antes de proceder a uma utilização definitiva do sistema de anotação, é preciso determinar sua executabilidade e a clareza do manual, o que é feito anotando-se um fragmento do material selecionado no passo 1. Preferencialmente, é interessante que pelo menos duas pessoas façam essa anotação paralelamente.

4. **Avaliação do teste**. Envolve comparar as decisões dos anotadores, além de decidir pelas medidas apropriadas sobre concordância na anotação e sobre como aplicá-las (ver, para este fim, Cohen, 1960; Krippendorff, 2004; Artstein e Poesio, 2008). É preciso ter em mente que o que se busca são anotações *confiáveis*, isto é, que sejam *estáveis* e *reproduzíveis*. A estabilidade está relacionada à concordância intra-anotadores, isto é, ao quanto um mesmo anotador é consistente na anotação. Já a reprodutibilidade está relacionada à concordância entre anotadores, isto é, ao quanto os anotadores concordam na anotação dos mesmos fenômenos. Estabilidade e reprodutibilidade são fundamentais para que o treinamento de algoritmos de aprendizagem de máquina seja eficiente. É preciso estabelecer o nível (mínimo) satisfatório de concordância entre anotadores. Enquanto o teste de anotação não atingir o mínimo satisfatório, volta-se ao passo 2 para redefinir o sistema de anotação e o manual.

5. **Anotação manual de grande parte do material**. Uma vez determinado que o sistema de anotação é satisfatoriamente executável, passa-se à anotação de grande parte do material, processo que pode levar meses ou anos.

6. **Treinamento de um analisador**. Quando for acumulada uma certa quantidade de material anotado manualmente (p.e., 100 mil palavras), pode-se começar a avaliar se um analisador pode ser eficientemente treinado para que a anotação semiautomática possa começar. Para isso, este material acumulado deve ser dividido em duas partes, uma para treinamento (p.e., 90% do material) e outra para teste, de modo que o analisador seja treinado com base na porção de treinamento e testado sobre a porção "inédita" de teste. Uma vez que para a porção teste há uma anotação manual correta (em princípio) disponível, o desempenho do analisador pode ser avaliado, quanto à acurácia em relação à anotação alvo (por exemplo, usando a medida PARSEVAL, cf. Black et al, 1991).

7. **Anotação semiautomática**. Se o desempenho do analisador se mostrar satisfatório, passa-se a utilizá-lo em novo material, em conjunto com a correção/revisão por anotadores humanos. Caso não seja satisfatório, podem ser necessários ajustes no sistema de anotação ou pode ser necessário mais material para treinamento. No primeiro caso, pode ser preciso voltar ao passo 2, a depender do quanto o sistema precisará ser alterado (em alguns casos, como se vê na seção 4, certas modificações podem ser aplicadas automaticamente sobre o *corpus*, não requerendo revisão manual). No segundo caso, volta-se ao passo 5, para produção manual de mais material anotado.

À medida que se avança na quantidade de material anotado semiautomaticamente, torna-se cada vez menos viável fazer modificações no sistema de anotação que demandem revisão manual do mesmo. Daí a importância de uma atitude criteriosa e cuidadosa nos passos 1 a 4, para que a necessidade de alterações eventualmente detectadas em etapas seguintes seja mínima.

3. A IMPORTÂNCIA DOS BANCOS DE ÁRVORES NA PESQUISA EM LINGUÍSTICA

Corpora anotados são importantes em todos os ramos da linguística, uma vez que constituem bases de dados perenes sobre as quais se podem efetuar análises qualitativas e quantitativas de vários tipos, que complementam outras abordagens como o recurso à intuição dos falantes ou ainda estudos baseados em experimentos, prática corrente em aquisição da linguagem e cada vez mais em análises sintáticas. Em linguística histórica, uma vez que não há falantes nativos disponíveis, os *corpora* são indispensáveis. Eles podem até abranger a totalidade dos dados disponíveis, quando se consideram os períodos mais antigos das línguas. A anotação morfossintática permite explorar de maneira consistente e reprodutível quantidades de dados inacessíveis ao trabalho manual, permitindo um acesso cada vez mais completo e confiável aos dados do passado.

Não só os fenômenos frequentes podem assim ser apreendidos de modo mais rigorosos, mas os fenômenos raros, que são, em certos casos, essenciais para a verificação de hipóteses sofisticadas sobre a natureza das gramáticas subjacentes aos textos, podem ser detectados e trazidos à tona por buscas automáticas complexas que escaneiam em segundos milhares de frases. Daremos aqui um rápido exemplo desse tipo. As línguas românicas e germânicas se diferenciam pela posição do advérbio nas orações em que o sujeito segue o verbo. Nas primeiras, o advérbio segue o sujeito posposto, nas segundas, o precede (Belletti 2004). Isso se deve à posição diferente ocupada pelo verbo nos dois tipos de línguas.

Nas línguas germânicas, o verbo precede o sujeito porque foi movido para uma posição mais alta, nas línguas românicas, o verbo não se moveu, e o sujeito posposto ocupa uma posição baixa, à direita do advérbio. As orações em que o sujeito é posposto e há um advérbio de modo são, portanto, essenciais para descobrir se uma língua tem uma sintaxe de tipo românico ou de tipo germânico, questão crucial para as línguas românicas antigas, em particular o português clássico. Mas tais orações (exemplificadas por (1) a seguir) são relativamente raras, e difíceis de achar manualmente em centenas de milhares de palavras.

(1) Em muytas occasiões advirtio *Deos* à Madre Elena **interiormente** o que convinha à sua honra, (C_002,196.810)

Essa dificuldade não existe para o programa *Corpus Search*, que em menos de um minuto é capaz de achar todas as frases correspondendo à descrição formulada na seguinte busca:

```
query: (tns_vb2 HasSister ADVP*) AND (ADVP* iDomsOnly ADV)
AND (ADV iDominates *mente) AND (tns_vb2 HasSister NP-SBJ*)
AND (tns_vb2 precedes NP-SBJ*) AND (NP-SBJ* precedes ADVP*)
```

Com base nas relações "HasSister" (é irmão), "iDominates" (domina imediatamente) e "precedes" (precede)[5], aplicando-se às categorias sintáticas ADVP (sintagma adverbial), NP-SBJ (sintagma nominal sujeito) e tns_vb2, (um conjunto definido a partir de determinadas categorias verbais), *Corpus Search* extrai (1) do *corpus*, bem como todas as frases análogas, produzindo a seguinte saída, em que cada nó tem um índice numérico[6]:

[5] Cf. http://corpussearch.sourceforge.net/CS-manual/Revise.html

[6] Note-se que a numeração dos nós automaticamente atribuída pela ferramente de busca é totalmente arbitrária, podendo exibir lacunas na sequência, sem que isso tenha relevância para os resultados.

```
/*
1 IP-MAT:  10 VB-D, 25 ADVP, 26 ADV, 27 interiormente, 12 NP-SBJ
*/
(0
  (1 IP-MAT
    (2 PP (3 P Em)
          (5 NP (6 Q-F-P muytas) (8 N-P occasiões)))
    (10 VB-D advirtio)
    (12 NP-SBJ (13 NPR Deos))
    (15 PP-ACC (16 P a@)
               (18 NP (19 D-F @a) (21 NPR Madre) (23 NPR Elena)))
    (25 ADVP (26 ADV interiormente))
    (28 CP-QUE (29 WNP-1 (30 D o) (32 WPRO que))
               (34 IP-SUB
                   (35 NP-SBJ *T*-1)
                   (37 VB-D convinha)
                   (39 PP (40 P a@)
                          (42 NP (43 D-F @a) (45 PRO$-F sua)
                                 (47 N honra)))))
    (49 , ,))
    (51 ID C_002,196.810))
/*
  source file, hits/tokens/total
  c_002_psd.txt 1/1/1272
*/
```

Um outro exemplo do papel que a construção de bancos de árvores desempenha em linguística se encontra na sua aplicação a línguas para as quais não existe uma longa tradição gramatical como é o caso das línguas indígenas sem cultura escrita. Nesse caso, a construção do sistema de anotação exige que se definam categorias de análise, a serem testadas na aplicação das ferramentas automáticas. Nesse caso, os etiquetadores e analisadores exercem uma verdadeira função heurística no sentido de que permitem testar hipóteses sobre fenômenos linguísticos (Lavid, 2013). Experiências desse tipo estão sendo realizadas no âmbito da plataforma *Tycho Brahe*[7], onde estão sendo construídos *corpora* anotados de línguas indígenas brasileiras e argentinas da família guaikuru, com os mesmos recursos computacionais usados para o CTB.

[7] Cf. http://www.tycho.iel.unicamp.br/tbf/login

4. PROCESSAMENTO AUTOMÁTICO DE TEXTOS

Agora que temos uma visão geral do que são bancos de árvores, de como são construídos e de sua importância para a pesquisa em linguística, discutimos nesta seção alguns aspectos de seu processamento automático. Destacamos a seguir uma parte fundamental da construção de bancos de árvores, a saber, a tarefa de análise sintática automática. Inúmeros estudos e métodos de análise automática têm sido desenvolvidos ao longo dos últimos vinte anos. O objetivo central aqui é o de demonstrar a relação intrínseca entre escolhas envolvendo sistemas de anotação sintática e a qualidade dos resultados obtidos nas tarefas de análise automática, no intuito de identificar diretrizes para a especificação de sistemas de anotação que contribuam para o alcance de melhores práticas na construção de bancos de árvores.

4.1. A anotação e o desempenho do analisador

A análise automática de uma sentença consiste em atribuir uma ou mais estruturas sintáticas a ela, de modo que as relações entre as palavras sejam explicitadas seja pela delimitação dos constituintes sintáticos que elas formam, seja pela identificação da função sintática dos elementos. Os analisadores modernos são, de modo geral, total ou parcialmente probabilísticos, isto é, produzem análises possíveis (quando mais que uma) de uma sentença e as ordenam conforme a probabilidade de cada uma. A análise com maior probabilidade é geralmente tida como a "melhor" (i.e., provavelmente mais correta) análise para uma dada sentença.

A "aprendizagem" de um analisador consiste, portanto, em receber exemplos de análise e construir um modelo probabilístico que lhe permita aplicar análises sobre novas sentenças, inclusive aquelas estritamente inéditas, isto é, cuja estrutura como um todo nunca tenha sido vista no *corpus* de treinamento. Isso é possível, porque o modelo probabilístico é construído como um vasto sistema de regras que se aplicam de modo local, isto é, isolando estruturas sintáticas que podem ocorrer em diferentes partes da sentença, em função do caráter combinatório da sintaxe.

E aqui fica mais evidente a importância da consistência na anotação: para que a aprendizagem de máquina seja capaz de produzir bons modelos, é necessário que um dado fenômeno, por exemplo, uma oração relativa, tenha sempre o mesmo tipo de análise sintática. Caso contrário, o modelo gerado no treinamento iria conter as várias análises inconsistentes para orações relativas, o que levaria o analisador a produzir também análises inconsistentes. Por outro lado, sabemos que as línguas produzem expressões ambíguas, isto é, que podem ter mais que uma interpretação. A sentença na Figura 1, inclusive, é um exemplo de expressão ambígua: o advérbio "hoje" pode tanto ser modificador do auxiliar "está", quanto do particípio "agendada". No primeiro caso, significaria que o agendamento é válido "hoje", mas não se sabe em que dia a audição ocorrerá de fato. No segundo caso, certamente a leitura preferencial, a audição é que ocorrerá "hoje".

Pode-se dizer que a consistência na aplicação do sistema de anotação é uma consistência *externa*, no que diz respeito às propriedades do sistema de anotação em si. Como discutido na seção 2, ela é fundamental para a construção de um *corpus* extenso, não apenas por garantir a executabilidade manual da anotação, mas também por permitir uma aprendizagem de máquina eficiente. Mas há uma consistência *interna* ao sistema de anotação, que diz respeito a como o sistema codifica os diversos tipos de informação (categoria e função sintática, classe morfológica, informações flexionais etc.). No restante dessa subseção, um estudo de caso visa a demonstrar a importância de pensar a consistência interna da anotação, com vistas à melhorar o desempenho da análise automática.

4.2. A revisão da anotação de verbos no CTB

A anotação sintática pode ser vista como a imposição de uma "camada sintática" sobre a "camada morfológica" que, no caso do CTB, consiste de tokens (palavras e pontuação) e suas respectivas etiquetas morfológicas.

Figura 2. Sentença etiquetada.

A Figura 2 exibe a versão simplesmente etiquetada da sentença da Figura 1. Em termos cronológicos da construção de bancos de árvores, a definição do sistema de etiquetas morfológicas em geral precede (às vezes, em muitos anos) a definição do sistema de anotação sintática (que passa a incluir o primeiro). Isso pode gerar certos desalinhamentos entre os dois sistemas, o que pode ter um impacto negativo significativo para a análise automática.

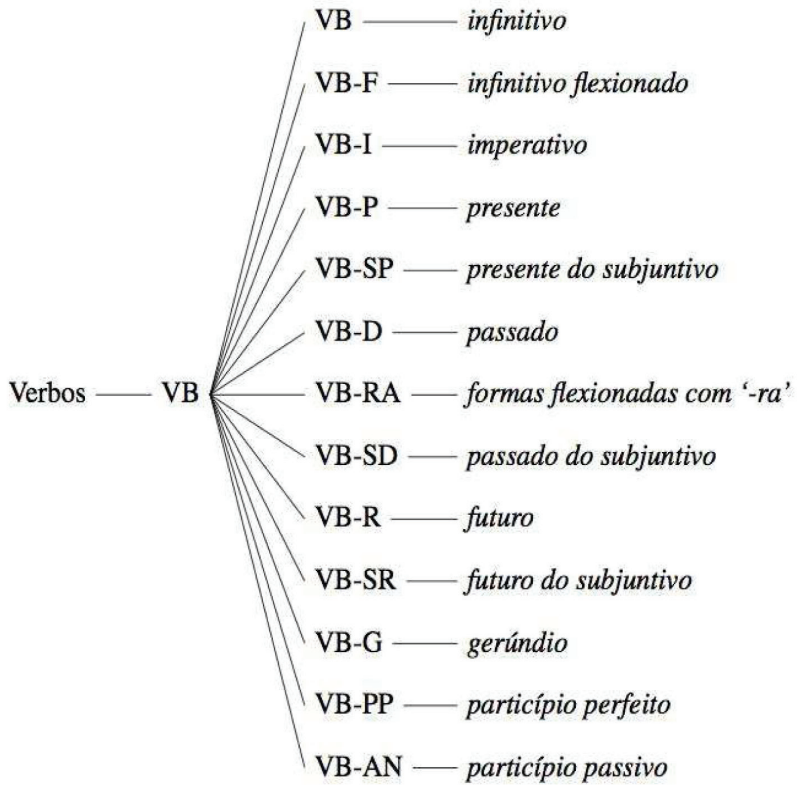

Figura 3. Etiquetas morfológicas para verbos em geral, no CTB.

Tomemos, como exemplo de caso, a anotação de sintagmas verbais no CTB. Vamos considerar, inicialmente, o atual sistema de etiquetas morfológicas, exibido na Figura 3. No CTB, verbos em geral recebem a etiqueta "VB" acrescida das subetiquetas pertinentes, exceto no caso de infinitivos que não recebem nenhuma. Além desta, há quatro classes especiais de verbos, para os verbos *ser, estar, haver e ter*, cujas etiquetas base são, respectivamente, SR, ET, HV e TR, e cujas subetiquetas seguem basicamente o mesmo sistema representado na figura. Neste sistema, todas as etiquetas verbais compartilham de uma mesma etiqueta "base", VB (SR, ET, HV ou TR). Quando a estrutura sintática entra em cena, são as subetiquetas (inclusive a subetiqueta vazia, no caso de infinitivos) que determinam as projeções sintáticas possíveis (previstas pelo sistema de anotação).

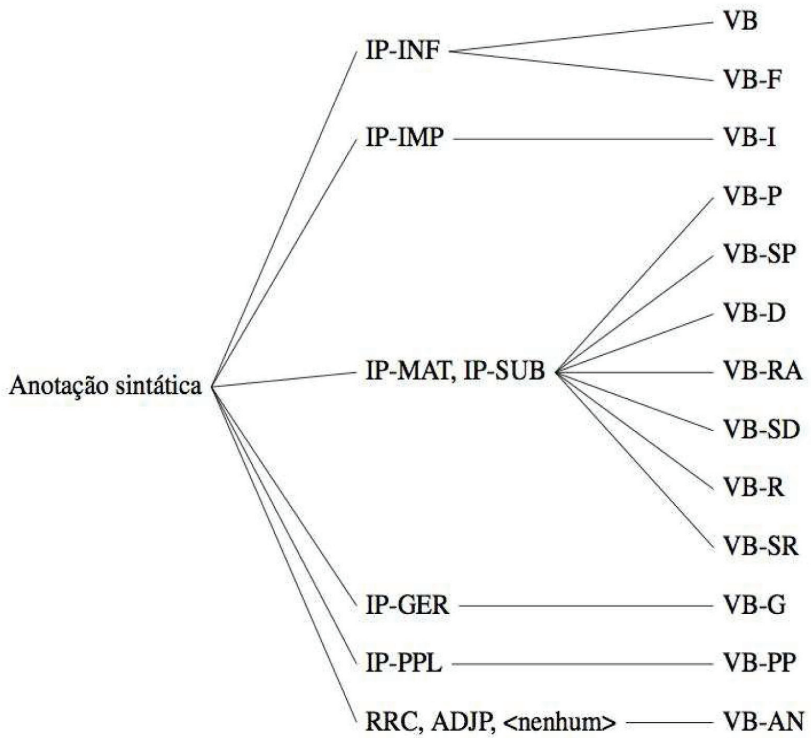

Figura 4. Projeções sintáticas possíveis para cada etiqueta verbal do CTB.

A Figura 4 exemplifica os vários tipos de projeção sintática possíveis, a depender do tipo de subetiqueta. Note-se que as treze etiquetas verbais previstas formam seis subgrupos, quando se consideram os tipos de projeção sintática possíveis e adequados a cada uma. As seis classes são: orações infinitivas (IP-INF), orações imperativas (IP-IMP), orações matrizes e subordinadas (IP-MAT e IP-SUB), orações gerundivas (IP-GER), orações participiais adverbiais (IP-PPL) e, por fim, a classe que varia conforme o contexto sintático em que ocorre a etiqueta VB-AN, configurando ora orações relativas reduzidas (RRC), ora sintagmas adjetivais, ora sem projeção alguma. Isso significa que, para modelar satisfatoriamente todas as possíveis combinações entre constituintes sintáticos e etiquetas morfológicas da Figura 4, o algoritmo de aprendizagem terá que atribuir (*grosso modo*) uma probabilidade para cada combinação.

Por exemplo, dada a configuração acima, o modelo probabilístico conterá duas regras para produzir um IP-INF e sete regras para produzir um IP-MAT ou IP-SUB. Tais regras terão probabilidades mais baixas fazendo com que a probabilidade geral das árvores também seja mais baixa. Isso aumenta as chances de que a melhor análise seja preterida por uma análise incorreta, em função das diferenças muito pequenas entre as probabilidades das análises concorrentes. Se, alternativamente, as sete regras para IP-MAT/SUB pudessem ser reduzidas a uma só regra, as chances da melhor análise prevalecer seriam maiores, por hipótese.

Para averiguar esta hipótese, foi concebido um sistema de etiquetas verbais alternativo. Neste, as subetiquetas -I, -G, -PP e -AN foram fundidas à etiqueta base. VB passa a fazer referência aos infinitivos, enquanto VBT aos verbos flexionados. A mesma revisão foi aplicada também para os verbos leves (etiquetas SR, ET, HV e TR), mencionados anteriormente. Neste sistema revisto (exibido na Figura 5, abaixo), a relação entre os constituintes sintáticos e as etiquetas base é agora de um para um. Para que nenhuma informação seja perdida, as subetiquetas são mantidas no sistema, porém não são fornecidas ao analisador, seja no treinamento, seja no teste de análise.

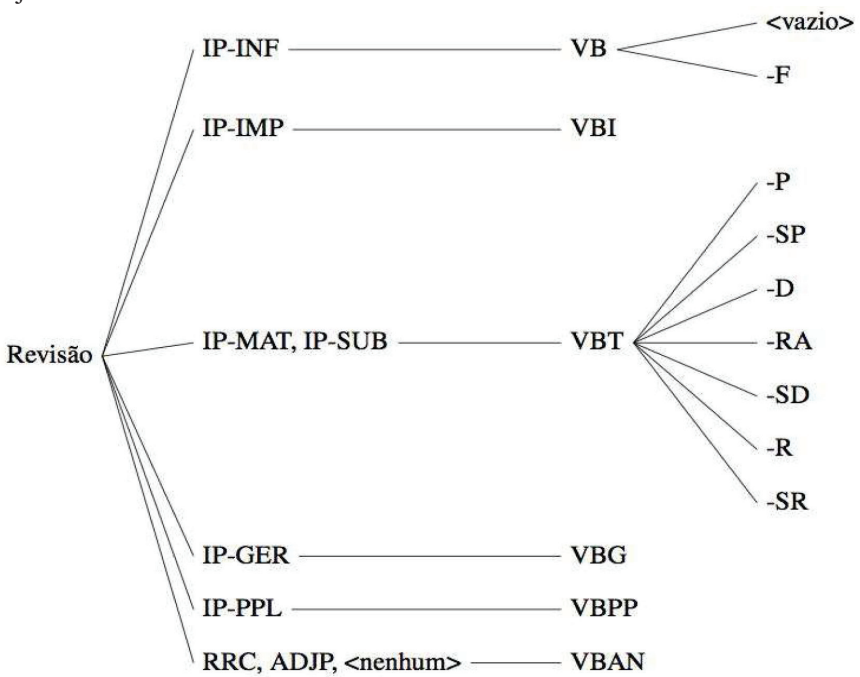

Figura 5. O sistema revisto de etiquetas verbais no CTB.

4.3. O teste do sistema revisado

O experimento foi conduzido com o analisador de Bikel (2004). Foi utilizada a versão de 2010 do CTB (Galves & Faria, 2010), que consiste em 16 textos de diversos autores e épocas, num total de 26.732 árvores sintáticas, sendo 556.899 tokens. Deste total, 1000 árvores (22.261 tokens) foram aleatoriamente extraídas para formar a seção de teste, de modo a evitar um enviesamento histórico (data de produção), de gênero ou de autoria dos dados. As 25.732 árvores restantes (534.638 tokens) foram utilizadas para treinamento. Das 1000 sentenças da seção de teste, 862 tem comprimento menor ou igual a 40 tokens. As seções de treinamento e teste inicialmente produzidas são as que configuram a condição "atual" descrita abaixo. A condição revisada é gerada a partir dessa.

Tabela 1. Condições experimentais.

Condição		Etiquetas	Descrição
1	Atual	305	Versão com sistema de anotação atual
2	Revisada	275	Sistema revisado e sem subetiquetas verbais

Para os experimentos, a partir de cada seção de teste, é extraída uma versão etiquetada correspondente, excluindo a estrutura sintática. A versão etiquetada é, então, submetida ao analisador e a análise produzida é comparada com a seção de teste original (chamada, neste caso, de "padrão-ouro" de anotação). Dessa forma, é possível estabelecer a acurácia do analisador usando a medida PARSEVAL (Abney et al., 1991). Duas condições básicas foram comparadas:

- Condição **atual**. O *corpus* de treinamento nessa condição mantém todas as subetiquetas, sintáticas e morfológicas. Para garantir que o analisador considere as subetiquetas, todas as etiquetas do sistema foram modificadas, trocando-se o "-" (hífen) que separa a base do restante pelo símbolo "_" (sublinhado), como em IP_INF, por exemplo. Com isso, o analisador trata toda a etiqueta como um símbolo atômico, isto é, indivisível. O número total de rótulos apresentados ao analisador no treinamento foi de 305 (somando etiquetas sintáticas e morfológicas).

- Condição **revisada**. Sistema de etiquetas verbais modificado, conforme discutido na seção anterior, e com subetiquetas verbais removidas, visto que por hipótese se tornam irrelevantes para o analisador após a mudança nas etiquetas base. Com esta mudança, o número total de rótulos apresentados ao analisador no treinamento foi de 275, ou seja, 30 rótulos a menos do que na condição atual.

Tabela 2. Desempenho médio do analisador para cada condição experimental, conforme calculado pelo aplicativo *evalb*.

Condição		F1	F1 (<=40)	F1 (pond.)	Parênt. cruzados	Sentenças sem análise[8]
2	Revisada	76,7856	79,3106	82,25011	1,23	13 (69-197)
1	Atual	74,5431	77,3538	80,23044	1,43	12 (69-197)

Os resultados do analisador são exibidos na Tabela 2 (ordem decrescente de desempenho). Para cálculo da precisão e da cobertura, foi utilizado o aplicativo *evalb*[9] (Sekine & Collins, 2013). Foi calculada a média harmônica F1 ("F-score" = 2*P*R / (P+R), sendo P a precisão e R a cobertura) entre precisão e cobertura para cada sentença e depois calculada a média geral da F1 para cada condição.

[8] Nos experimentos, um tempo limite foi especificado para o processamento de cada sentença, de modo que o analisador "desistia" da sentença, se não conseguisse produzir uma análise em até 10 minutos. Tais análises demoradas tendem a ser demasiadamente precárias, razão pela qual não vale a pena alongar o processamento como um todo em função dessas poucas sentenças.

[9] *Software* amplamente utilizado para avaliação de análises sintáticas alternativas em que uma árvore candidata é comparara a uma árvore alvo quanto ao percentual de constituintes em comum.

O mesmo cálculo foi feito apenas para as sentenças com comprimento menor ou igual a 40 tokens, que compõem 86,2% do *corpus* de teste. Por fim, visto que foi detectada uma correlação moderada significativa[10] entre o comprimento das sentenças e a F1, foi calculado o desempenho ponderado pelo comprimento das sentenças. É possível que esta medida seja mais realista quanto ao desempenho do analisador. A duas últimas colunas informam, respectivamente, o número médio de parênteses cruzados e o número de sentenças que ficaram sem análise (ver nota) juntamente com os comprimentos mínimo e máximo de tais sentenças.

4.4. Discussão parcial dos resultados

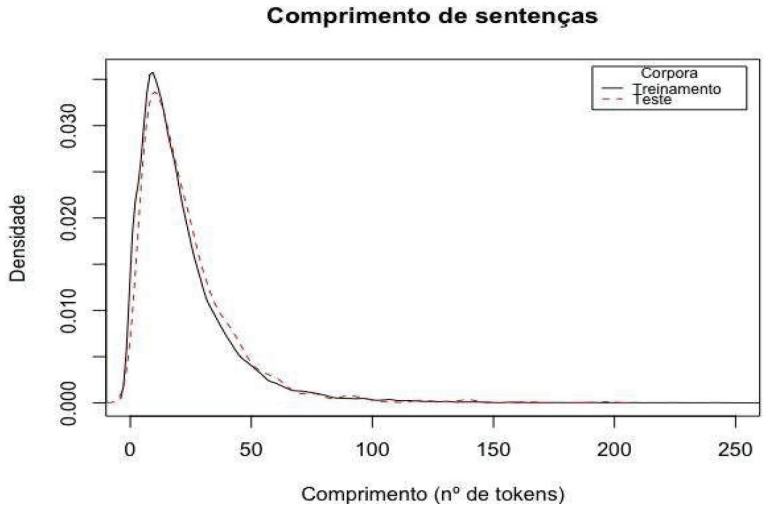

Gráfico 1. Distribuição nos *corpora* de treinamento e teste (comprimento das sentenças).

Como questão preliminar, seria a seção de teste suficientemente representativa do *corpus* para que os resultados obtidos sejam indicativos confiáveis do desempenho para novos textos? Supondo que o material inédito seja similar (em estilo e gênero) ao material que compõe o *corpus* de treinamento, o Gráfico 1 indica que sim. Neste, vemos uma distribuição similar para as duas seções, em termos do comprimento das sentenças.

Dito isto, os resultados apresentados na Tabela 2 confirmam a hipótese inicial de que a alteração no subsistema de etiquetas verbais tornaria o analisador mais eficiente. Em termos numéricos, a condição 2 obteve uma melhora de aproximadamente dois pontos percentuais quando comparada à condição 1, um ganho que justifica tornar esta modificação definitiva, uma vez que pode ser implementada automaticamente e não implica em perda de informação na anotação. A F1 (<=40) de 79,31 obtida na condição 2 indica que, em termos comparativos, o desempenho da análise para o CTB a aproxima de resultados para outras línguas, como mostra a Tabela 3.

[10] As correlações para as condições experimentais variaram de -0.475357 a -0.590796 ($p < 2.2e\text{-}16$).

Tabela 3. Alguns resultados de análise automática encontrados na literatura e o do experimento realizado.

Banco de árvores	Língua	F1	Treinamento
WSJ + NANC (McClosky et al., 2006)	Inglês	92,1	40000 (WSJ) + 1750 (NANC)
Tüba-D/Z (Klübler et al., 2008)	Alemão	88,5 (<=40)	25005 sentenças
CTB	Português	79,31 (<=40)	25732 sentenças
TIGER (Klübler et al., 2008)	Alemão	77,3 (<=40)	25005 sentenças

Os resultados ainda estão relativamente distantes dos melhores obtidos para o inglês sobre o Penn Treebank, como vemos na Tabela 3. No entanto, é esperado que os resultados para o inglês sejam melhores, dado que os analisadores são em sua maioria desenvolvidos com base no *corpus Wall Street Journal* (WSJ) e o sistema de anotação deste *corpus* é mais simples. Por outro lado, é possível que haja espaço para maiores avanços na análise do CTB, em particular, através de outras revisões do sistema de anotação, melhoria da qualidade e acréscimo de material de treinamento. Faz-se necessária, ainda, a avaliação de outros analisadores disponíveis, como o de McClosky et al. (2006), por exemplo.

5. A ESPECIFICAÇÃO DO SISTEMA DE ANOTAÇÃO E O TREINAMENTO DO ANALISADOR

Os resultados apresentados na seção anterior nos permitem elencar dois princípios norteadores que ensejam boas práticas na especificação do sistema de anotação e no treinamento de analisadores, aspectos importantes do fluxo de construção de bancos de árvores de estrutura sintagmática em que rótulos sintáticos e etiquetas fazem uso da distinção base/subetiqueta, tais como o CTB. O primeiro princípio pode ser descrito como em (2):

(2) A base das etiquetas deve codificar tão somente e exaustivamente as distinções relevantes para a análise sintática.

Ser relevante para a análise sintática significa implicar ou na projeção de uma categoria sintática ou na determinação de uma subetiqueta sintática ou ambas as coisas. O intuito deste princípio é o de incentivar a especificação de um sistema conciso de etiquetas que seja, ao mesmo tempo, sintaticamente consistente e informativo. O mesmo princípio deve guiar também a definição dos rótulos sintáticos e das relações entre categorias sintáticas. Uma vez definido o sistema

de anotação, deve-se preparar um *corpus* de treinamento sem as subetiquetas morfológicas, visto que um sistema mais conciso tenderá a produzir melhores resultados. Isso é possível se houver uma versão prévia etiquetada do texto a ser analisado. Isso nos leva ao princípio em (3):

(3) O *corpus* de treinamento do analisador deve excluir toda informação supérflua para a análise sintática, desde que esta seja automaticamente recuperável.

Vale ressaltar que o princípio acima, bem como quaisquer outras decisões envolvendo o analisador dependem, fundamentalmente, de um conhecimento adequado de seu funcionamento e de sua configuração. Um analisador mal configurado ou operando em modo "genérico" (i.e., sem predisposição para particularidades da língua) certamente terá um desempenho bastante limitado, quando comparado ao estado da arte.

6. CONSIDERAÇÕES FINAIS

O estudo apresentado neste artigo tinha por objetivo verificar a hipótese de que a revisão do subsistema verbal de etiquetas morfológicas do CTB melhoraria a desempenho do analisador. Essa melhoria seria consequência de um sistema mais informativo e menos redundante de etiquetas. Experimentos com o analisador foram conduzidos e avaliados para comparar diferentes condições de treinamento, em particular, comparar o sistema atual ao sistema revisado. Os resultados mostram um ganho aproximado de dois pontos percentuais, passando de 77,35% para 79,31% (para sentenças com até 40 tokens), um resultado que justifica adotar o sistema revisado.

O presente estudo é parte de uma iniciativa mais ampla que visa aumentar a qualidade de bancos de árvores e melhorar a produtividade na sua construção. Portanto, estão em andamento estudos sobre o impacto de outras intervenções no sistema de anotação, estudos comparativos para avaliar o desempenho de diferentes analisadores, e estudos para desenvolvimento de métodos de detecção de inconsistências e erros de anotação em bancos de árvores. Espera-se que, em conjunto, tais estudos resultem em sugestões concretas de boas práticas na construção de bancos de árvores, bem como na disponibilização de mais ferramentas computacionais para sua construção, manutenção e revisão.

REFERÊNCIAS BIBLIOGRÁFICAS

ABNEY, S., S. Flickenger, C. Gdaniec, C. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. (1991). Procedure for quantitatively comparing the syntactic coverage of english grammars. In E. Black, editor, *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 306–311, Stroudsburg, PA, USA. Association for Computational Linguistics.

ARTSTEIN, Ron e Massimo Poesio. (2008). Inter-coder agreement for Computational Linguistics (survey article). *Computational Linguistics*, 34/4, 555-596.

BIKEL, Dan. (2004). Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4).

BLACK, E., S. Abney, S. Flickenger, C. Gdaniec, C. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, T. Strzalkowski. (1991). Procedure for quantitatively comparing the syntactic coverage of English grammars, *Proceedings of the workshop on Speech and Natural Language*, p.306-311, February 19-22, Pacific Grove, California.

CHE, Wanxiang, Zhenghua Li, and Ting Liu. (2012. *Chinese Dependency Treebank 1.0* LDC2012T05. Web Download. Philadelphia: Linguistic Data Consortium.

COHEN, Jacob. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measument* 20: 37-46.

GALVES, Charlotte, e Pablo Faria. (2010). *Corpus Histórico do Português Tycho Brahe*. URL: http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html, acessado em 09/10/2016.

KRIPPENDORFF, Klaus. (2004). Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411-433.

KÜBLER, Sandra, Wolfgang Maier, Ines Rehbein e Yannick Versley. (2008, May). How to Compare Treebanks. In LREC.

LAVID, Julia. (2013). The Impact of Corpus Annotation on Linguistic Research: Theoretical and Methodological Challenges. In: ARIAS, Rosario, Mirian L. Rodríguez, Antonio M. Ortiz & Chantal P. Hernández. *Hopes and Fears: English and American Studies in Spain. Proceedings of the 36th AEDEAN Conference.* Dpto. de Filología Inglesa, Francesa y Alemana, Universidade de Málaga.

LEECH, Geoffrey N. (2009). *An Academic Autobiography*. URL: http://www.lancaster.ac.uk/fass/doc_library/linguistics/leechg/Autobiog.pdf, acessado em 08/06/2016.

MARCUS, Mitchell P., Mary Ann Marcinkiewicz e Beatrice Santorini. (1993). Building a large annotated corpus of English: the penn treebank. *Comput. Linguist*. 19, 2 (June 1993), 313-330.

MCCLOSKY, David, Eugene Charniak e Mark Johnson. (2006). Effective Self-Training for Parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference,* June, New York City, USA, Association for Computational Linguistics, p. 152-159.

REHBEIN, Ines e van Genabith, Josef. (2007). Why is it so difficult to compare treebanks? TIGER and TüBa-D/Z revisited. In: *TLT 2007 - The 6th International Workshop on Treebanks and Linguistic Theories*, 7-8 December, 2007, Bergen, Norway.

SEKINE, Satoshi e Michael John Collins. (2013). *Evalb software*. Disponível na internet em http://nlp.cs.nyu.edu/evalb/.

TAYLOR, Ann,Mitchell Marcus, e Beatrice Santorini. (2003). The Penn treebank: an overview. In: *Treebanks*. Springer Netherlands. p. 5-22.

XIAO, R. Z. (2008). Well-known and influential corpora. In A. Ludeling, & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook*. (Vol. 1). (Handbooks of Linguistics and Communication Science). Berlin: Mouton de Gruyter.