Araucárias: um corpus de árvores sintáticas para o ensino da sintaxe do português brasileiro

Wagner Ferreira Lima
Departamento de Letras Vernáculas e Clássicas
Universidade Estadual de Londrina
Londrina, Paraná, Brasil
wflima@uel.br

Cinthyan R. S. C. de Barbosa
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina, Paraná, Brasil
cinthyan@uel.br

ABSTRACT

Automatic syntactic parsers can support the teaching of Brazilian Portuguese. However, existing corpora that underpin such parsers are primarily designed for industrial applications rather than for the instruction of the vernacular. In this context, we propose the Araucárias corpus, a collection of syntactic trees derived from well-formed Portuguese sentences. The aim is to demonstrate its relevance for educational purposes in comparison to other existing corpora. Due to its relatively intuitive, simple, and evocative nature—attributes that are absent in the evaluated existing corpora—Araucárias is expected to offer a more viable approach to teaching Portuguese syntax.

CCS CONCEPTS

• Applied computing • Education

KEYWORDS

Araucarias, corpus of syntactic trees, teaching Brazilian Portuguese.

1 Introdução

Ler e escrever bem é um requisito para o pleno exercício da cidadania. Porém, o desenvolvimento dessas habilidades requer um esforço cognitivo adicional, a saber, o entendimento e o uso do conhecimento metalinguístico. Por exemplo, uma redação inteligível começa pela construção de sentenças bem formadas, e escritas de acordo com a norma culta. Logo, o desenvolvimento de habilidades sintáticas, tais como analisar e produzir sentenças bem formadas, é o tópico central deste artigo.

Tecnologias de Processamento de Língua Natural (PLN), por meio de seus produtos (ferramentas, recursos e/ou aplicativos), podem otimizar o desenvolvimento de tais habilidades. A questão é se estão efetivamente fazendo isso, pois por princípio aplicações devem ser vistas como facilitadores de procedimentos. Tal facilitação deve transparecer em todos os níveis, a começar pelo corpus que serve de base para as aplicações.

Tendo isso em mente, assumem-se três condições às quais um corpus precisa satisfazer se o propósito é facilitar o entendimento das regras sintáticas: (*a*) ser relativamente intuitivo (as descrições devem ser relativamente isomórficas com a descrição escolar feita das orações); (*b*) ser relativamente simples (as anotações devem empregar etiquetas básicas sugestivas); (*c*) ser relativamente evocativo (as descrições devem lembrar as descrições sintáticas vistas na escola).

Como se pretende mostrar, alguns corpora falham em satisfazer esses critérios; não porque eles são precários, longe disso; mas sim porque o enfoque deles parece ser prioritariamente a indústria e não o ensino [1][2]. Diante disso, propõe-se o *Araucárias*[1], um corpus de árvores sintáticas bem formadas que atende às condições acima e *ipso facto* pode servir a aplicações de PLN voltadas para o ensino da sintaxe do português brasileiro.

Essas árvores são modeladas em termos da Gramática Gerativa (GG) clássica [3], a qual mantém algumas similaridades descritivas com a Gramática Tradicional (GT) [4] [5] sem, no entanto, incorrer nas inconsistências metodológicas desta última. Tais similaridades, como se pretende mostrar, são salutares para o ensino do vernáculo que se baseia nos conhecimentos da linguística.

Assim sendo, este trabalho visa apresentar as bases teóricas e metodológicas do *Araucárias* e argumentar que, em comparação com alguns corpora existentes, esse corpus se mostra como o mais adequado às atividades de ensino. Antes, porém, realizar-se-á uma breve introdução teórica do assunto e uma rápida revisão de trabalhos congêneres, à luz das condições ora assumidas.

A apresentação se organiza como se segue: em "Similaridades entre descrições sintáticas" (Seção 2), sugere-se que existe uma similaridade fraca entre o gerativismo e a gramática tradicional; em "Trabalhos correlatos" (Seção 3), discute-se brevemente os corpora de árvores sintáticas para o português; em "Metodologia" (Seção 4), descreve-se o pipeline para a criação do *Araucárias*; em "Discussão dos dados" (Seção 5), compararam-se brevemente as

[1] Araucária é um pinheiro típico do Sul e símbolo do estado do Paraná.

diferentes abordagens descritivas, retomando a definição das referidas condições; e em "Considerações finais" (Seção 6), conclui-se a discussão com um prognóstico da pesquisa.

2 **Similaridades entre descrições sintáticas**

Em PLN, e naturalmente em Linguística, a análise sintática (*parsing*) pode acontecer segundo dois princípios: dependência e constituência [3]. De acordo com o princípio da dependência, a unidade da oração é determinada pela dependência interna de suas partes, ou seja, das palavras que a compõem; tal que uma se conecta diretamente a outra, formando uma hierarquia encabeçada pelo verbo centro da oração. Vale lembrar que muitos corpora usados em PLN seguem em geral esse princípio (Figura 2) [1][2].

Já conforme o princípio da constituência, a economia da oração se deve à hierarquia de seus constituintes, tanto os símbolos terminais, ou seja, as palavras, quanto os símbolos não-terminais, isto é, as categorias sintáticas que medeiam a relação das palavras com o todo; de maneira que essas não se mostram conectadas diretamente entre si. Representante desse princípio é o gerativismo linguístico, que é base para a construção do *Araucárias*.

A Gramática Gerativa (GG) [3] nasce como uma abordagem que objetiva identificar as sentenças de uma língua, ou seja, as cadeias de símbolos terminais (palavras) geradas a partir de uma gramática. Logo, a gramática é concebida como um dispositivo gerador de sentenças. Esse dispositivo é descrito como um conjunto de regras de produção.

Embora essa abordagem tenha sofrido mudanças ao longo do tempo, a teoria X-barra [7] sendo uma das versões mais recentes do gerativismo, o *Araucárias* emprega a versão original do conjunto finito de regras a serem adaptadas ao nosso propósito (Figura 1) [3][8]. Isso se deve a que as regras do gerativismo guardam *mutatis mutandis* similaridades com as regras da Gramática Tradicional (GT), de modo a atender, com algumas adaptações, às condições descritas acima.

Alguns exemplos: A regra "O -> SN + SV" especifica uma estrutura de oração dicotômica, que mapeia a estrutura "sujeito + predicado" apregoada pela GT. Já a regra "SV -> V SN", ao declarar que o verbo é o núcleo do SV e que, junto com ele, pode ocorrer um SN, evoca a descrição sintática tradicional de que o "objeto direto", sendo complemento de verbo transitivo direto, está dentro do predicado verbal.

O mesmo acontece com "SV -> V SP_C, em que o sintagma preposicional complemento (SP_C), o qual coocorre com o verbo dentro de SV, lembra o comportamento do "complemento relativo" descrito pela GT [4] [5]: o complemento relativo, sendo complemento de verbo transitivo indireto, está dentro do predicado também.

Finalmente, a regra "SN - > Det N" declara que um nome substantivo (N) é o núcleo do sintagma nominal (SN) e que ele pode receber uma determinação nominal (Det). Essa descrição é similar àquela da GT para os grupos nominais (por exemplo, sujeito, objeto direto e objeto indireto), em que os grupos são descritos como tendo um substantivo por núcleo, o qual pode receber uma determinação nominal (por exemplo, um artigo, um adjetivo, um numeral etc.).

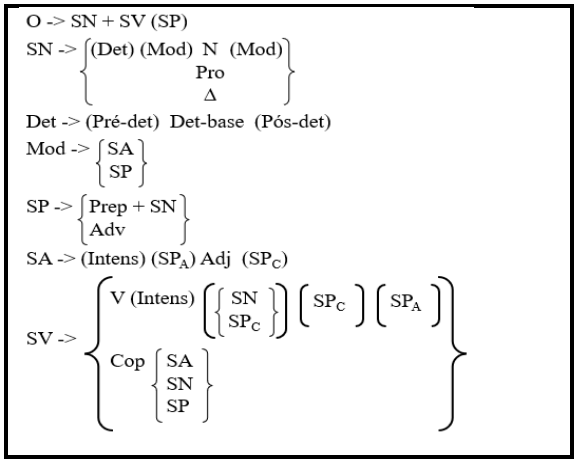

Figura 1: **Conjunto de regras de produção. Adaptado de Souza e Silva, e Koch [8]**

Por isso, pode-se dizer que a versão padrão do gerativismo é relativamente similar à concepção sintática preconizada pela GT; o que é desejável para uma abordagem científica e não disruptiva do ensino da sintaxe. Assim, com algumas adaptações, essa versão pode oferecer descrições isomórficas com as intuições sintáticas dos aprendizes e assegurar um aprendizado mais suave do vernáculo.

Embora esse entendimento expresse uma visão razoável da pedagogia do português, corpora há que, desprovidos de propósitos pedagógicos explícitos, apresentam descrições sintáticas arbitrárias e, assim sendo, dificultam o seu uso em atividades de ensino.

3 **Trabalhos correlatos**

Floresta Sintá(c)tica [2] e *Porttinari* [1] são dois corpora de árvores sintáticas (*treebanks*) bem estabelecidos e sofisticados. O *Floresta* foi lançado em 2000 e reeditado em 2007, com a adição de mais informações linguísticas. Inclusive, um subconjunto desse corpus, o *Bosque*, faz parte do pipeline do *Spacy* [9] para Língua Portuguesa (Figura 2). Já o *Porttinari 2.0*, por usa vez, é mais recente. Ele foi lançado em 2021 e recebeu uma atualização em 2023.

O mérito desses trabalhos é inquestionável. Com efeito, eles oferecem uma descrição das dependências entre as palavras da oração para qualquer sentença dada como entrada, uma vez que o modelo foi treinado sobre o corpus *Bosque*, que é um subgrupo do *Floresta Sintá(c)tica* [2]. Por exemplo, na descrição da oração "Os professores corrigem as provas.", pode-se visualizar as conexões

internas entre as palavras, representadas por setas que saem da palavra núcleo para a palavra subordinada (por exemplo, "professores" → "Os" ou "NOUN" → "DET") (Figura 2).

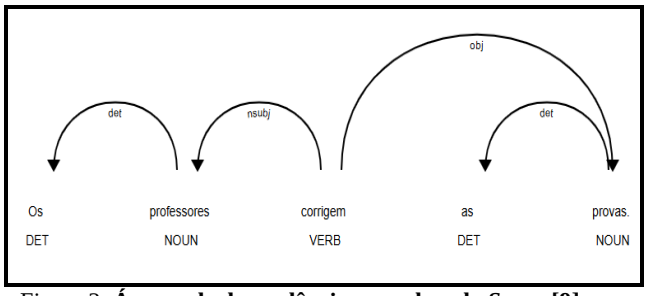

Figura 2: **Árvore de dependências gerada pelo *Spacy* [9] com base no *Bosque*.**

Não obstante a correção dessa análise à luz do princípio da dependência, ela nada sugere sobre a organização das palavras em categorias sintáticas integradoras. Por exemplo, "os" e "professores" formam juntos uma estrutura única, o "sujeito" da oração, que é a categoria integradora que se relaciona com o verbo "corrigem". A visualização dessa descrição é relativamente fácil, pois a oração é simples. Entretanto, com orações complexas, com muitos termos, o diagrama começa a ficar muito confuso e quase inteligível.

Essas constatações sugerem que esses corpora são incapazes de atender, nas situações de prática pedagógica, às três condições acima ((*a*) ser intuitivo; (*b*) ser simples; e (*c*) ser evocativo). As descrições do *Floresta* tanto quanto do *Porttinari* em (quase) nada refletem a intuição de orações como estruturas formadas por "sujeito" e "predicado" e, dentro desse, de "verbo" e "complementos" (não *a*).

É didático dizer que o "sujeito" é o assunto ou tema da oração enquanto o "predicado" é aquilo que se diz acerca desse assunto. Além disso, a conexão entre "verbo" e seus complementos ("objeto direto", "objeto indireto" etc.) é mais forte do que a relação do "verbo" com o "sujeito". Isso porque, de um lado, o complemento verbal é especificado pelo "verbo" e, também, ajuda a especificar o verbo no caso de ambiguidade; de outro, porque o "sujeito" não é específico de nenhum "verbo", estando presente em quase todas as orações.

Tampouco se pode alegar que esses corpora satisfazem a condição *b*. Com efeito, suas etiquetas estão longe de ser sugestivas e simples (não *b*). Por exemplo, os rótulos estão em inglês (o que não é a princípio problemático) e padronizados para serem universais. Aliás, esse é o propósito desses corpora. Além disso, as conexões, indicadas por setas (→), recebem elas mesmas determinações mais específicas ("nsubj", "det", "obj" etc.), relacionadas à função sintática das palavras; o que acaba por rebuscar o diagrama.

Por tudo que já foi dito, é pouco provável que esses corpora atendam à condição *c*. As descrições contidas nos diagramas de dependência dificilmente vão conseguir evocar experiências prévias com o ensino-aprendizado da sintaxe (não *c*). Tais experiências são úteis ao docente, quem precisa de uma base conceitual compartilhada para ancorar suas explanações. Assim, *Floresta* e *Porttinari* se mostram arbitrários e disruptivos.

Uma abordagem descritiva mais condizente com o ensino do vernáculo pode ser encontrada em Souza e Silva, e Koch (1995), as quais apresentam a teoria sintática gerativa de forma bastante didática. Isso se deve a que as autoras se apoiam na versão inicial do gerativismo [5], e não nas versões mais recentes, e implementam descrições de orações bem formadas e estruturalmente simples.

Com isso, elas produziram uma gramática do português escrito que é de uma só vez científica e evocativa. Científica porque se baseia nos princípios de descrição da GG [3][8]; evocativa porquanto não rompe radicalmente com a tradição gramatical. As árvores do *Araucárias* refletem, dessa maneira, as estruturas usadas por essas autoras para explicar a sintaxe da língua.

Mesmo assim, tal aplicação não satisfaz integralmente a condição *b*, pois a nomenclatura utilizada é *ipsis litteris* a da sintaxe gerativa, que é um tanto arbitrária. Por essa razão, alguns ajustes terminológicos foram necessários. Basicamente, substituiu-se a etiqueta "Det" – usada para denotar seja pronomes adjetivos, quantificadores ou numerais –, por etiquetas morfossintáticas mais sugestivas: "Pron", "Quant" e "Num", respectivamente; e "N", usada para "nome", por "Sub", abreviação de "substantivo" e notação mais evocativa.

4 Metodologia

O pipeline para a construção do corpus é como se segue:

(a) Geração de dados sintáticos sintéticos e bem formados. Optou-se por exemplares escritos e estruturalmente simples, tais como "As flores enfeitam os jardins na primavera." [8] A princípio, serão consideradas apenas orações de período simples, pois compreensão da estrutura delas é a base para o entendimento das orações de período complexo, as orações subordinadas da GT. As condições de obtenção dos dados são totalmente controladas:

(1) foi priorizada a classe de frases que na GT é conhecida por "frase oracional" ou "oração", ou seja, frase que contém verbo. Isso porque é esse tipo de frase que é prioritariamente estudado pela GT no ensino médio. Assim, um prompt foi criado, e também testado, que solicita ao modelo de Inteligência Artificial Generativa (IAG) uma amostra de 300 orações de acordo com o padrão de estrutura sintática desejado. A princípio, um total de 20 padrões (n = 6.000 orações) será solicitado;

(2) esses padrões de estrutura foram selecionados para capturar um número significativo de estruturas sintáticas bem formadas, ou seja, estruturas representativas e informativas do português escrito [10];

(3) além disso, porque são gerados via Inteligência Artificial Generativa, eles também vão ser balanceados [10].

(b) *Anotação dos dados*. Foi criado um programa em Python para extrair as árvores sintáticas desejadas: a classe *CriarDicionario( )* processa os dados de entrada, desfazendo, por exemplo as contrações existentes ("do" > "de o", "pelo" > "por o" etc.), e solicita o preenchimento das categorias morfossintáticas existentes ("artigo", "substantivo", "verbo" etc.).

O preenchimento das categorias encontradas na oração de entrada é realizado pelo método *preencher_categorias( )*, o qual devolve um dicionário contendo as palavras dessa oração. Por exemplo, a oração "Os professores corrigem as provas durante a semana.", a qual contém o mesmo padrão sintático da oração aludida na etapa (a) deste pipeline, depois de anotada, é transformada no dicionário da Figura 3.

```
testagem.preencher_categorias()

Artigos definidos/indefinidos: os as a
Substantivos/nomes próprios: professores provas semana
Verbos principal/subordinado: corrigem
Preposições de adjunto da oração: durante

O dicionário resultante é:

{'artigo': ['os', 'as', 'a'],
 'quantificador': [],
 'numeral': [],
 'pronome-adjetivo': [],
 'pronome-pessoal': [],
 'adjetivo': [],
 'substantivo': ['professores', 'provas', 'semana'],
 'copula': [],
 'auxiliar': [],
 'verbo': ['corrigem'],
 'adverbio': [],
 'adverbio-intensidade': [],
 'prep-adjunto-oracao': ['durante'],
 'preposicao-adjunto': [],
 'preposicao-complemento': []}
```

Figura 3: **Note-se que esse método é chamado mediante a variável "testagem", a qual instanciou, neste caso, a classe *CriarDicionario( )*.**

Já a classe *AnaliseMorfossintatica( )* gera propriamente a árvore sintática, a qual juntamente com a oração original correspondente vai compor o corpus. A correção da árvore gerada pode ser verificada por meio do método de visualização *mostrar_arvore( )*[2] (Figura 4).

```
testagem_1.mostrar_arvore()
```

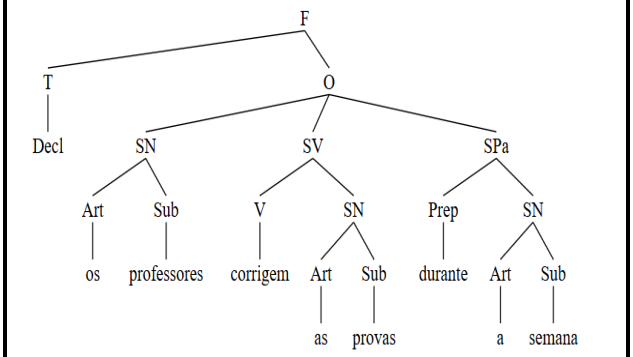

Figura 4: **Veja-se que esse método é chamado mediante a variável "testagem_1", a qual instanciou, neste caso, a classe *AnaliseMorfossintatica( )*.**

(c) *Armazenamento dos dados*. A função *armazenar_dados()* guarda provisoriamente os dados obtidos em uma lista, a qual mais tarde é salva em um arquivo .xlsx para posterior encaminhamento, como, por exemplo, criar um *dataset* para aprendizagem de máquina. Todas essas etapas já foram testadas (Figura 5).

```
oracao = sentenca
estrutura = testagem_1.editar_etiquetas()

dados = armazenar_dados(df=df, oracao=oracao, estrutura=estrutura)
dados
```

	Orações	Estruturas
0	As mães arrumam as casas pela manhã.	(F (T (Decl)) (O (SN (Art as) (Sub mães)) (S...
1	Os amigos assistem os filmes à noite.	(F (T (Decl)) (O (SN (Art os) (Sub amigos)) ...
2	Os professores corrigem as provas durante a se...	(F (T (Decl)) (O (SN (Art os) (Sub professore...

Figura 5: **Note-se que a função *armazenar_dados( )* recebe como parâmetros "df" (*dataframe* a ser preenchido); "oração" (sentença da planilha inicialmente gerada pelo modelo de IAG); e "estrutura" (a "árvore deitada" de cada oração).**

5 Discussão dos dados

Em comparação com o *Floresta* e o *Porttinari*, o *Araucárias* se apresenta como uma abordagem viável para o ensino da sintaxe do português brasileiro. Com efeito, ele atende a todos os critérios ora assumidos. Contém descrições que são relativamente intuitivas,

[2] F (Frase); T (Tipo de frase); O (Oração); SN (Sintagma nominal); SV (Sintagma verbal); SPa (Sintagma preposicional adjunto); Art (Artigo); Sub (Substantivo); Prep (Preposição).

notação simples e sugestiva e apelo a experiências escolares prévias. Senão, veja-se.

5.1 **Árvores contendo descrições intuitivas**

Como já antecipado, essa condição se caracteriza por uma isomorfia fraca entre as representações sintáticas da GG [3][8] e aquelas que os aprendizes adquiriram após passarem pelo ensino de português baseado na GT [3][4]. Intuição é uma noção psicológica complexa cuja definição ultrapassa o escopo deste artigo. Porém, neste trabalho, intuição pode ser *grosso modo* entendida tanto como a imageria mental visual das conexões entre os termos ou constituintes da oração, quanto como a disponibilidade mental para evocar similaridades entre a descrição sintática da GG e a descrição sintática da GT. (Este último é o sentido mais evidente que o conceito assume neste trabalho.)

E, quando se alega a existência de uma isomorfia, mesmo fraca, entre as duas formas de descrever a sintaxe das orações, as da GG e GT, está-se a afirmar que uma descrição ou representação sintática pode ser *mutatis mutandis* mapeada na outra. Um exemplo pode ajudar nesse sentido. Como já comentado, uma das regras da GG [3][8] declara a subordinação de SN e SV a O – categoria sintática mais abstrata e integradora de categorias menos abstratas e mais específicas.

Essa regra mapeia, portanto, a representação sintática apregoada na escola de que uma oração é formada por sujeito e predicado, seus termos básicos [4]. Assim, mesmo que os aprendizes não se recordem imediatamente dessas conexões, nem tampouco associem as duas descrições, a árvore sintática da GG é tão intuitivamente visual que pode, com algum esforço docente, evocar tais representações e relacionar uma com a outra.

Os diagramas de dependências que constituem os corpora *Floresta* e *Porttinari*, por outro lado, são também visuais e relativamente intuitivos (Figura 2). Com efeito, eles conseguem evidenciar a conexão direta entre as palavras que compõem a oração. Todavia, esse modo de descrever a oração, conquanto estruturalmente correto e conceitualmente plausível, é nada isomórfico com as representações da GT.

Por essa razão, trata-se de árvores pouco atraentes para o ensino da sintaxe do ponto de vista linguístico. Pode-se dizer, de acordo com isso, que elas se mostram disruptivas em relação às representações da GT. Assim, da maneira como se apresentam, elas podem dificultar, mais do que facilitar, a instrumentalização do conhecimento linguístico de sintaxe.

5.2 **Árvores representando estruturas simples**

Quem é professor sabe que a abordagem de qualquer assunto a ser ministrado deve começar pelos exemplos mais simples e evoluir para os casos menos simples ou complexos. Com o ensino da sintaxe, e da linguagem verbal em geral, isso não é diferente. Assim como acontece com definição de intuição, a definição de simplicidade tampouco é formalmente clara.

Para os efeitos deste projeto, entende-se por simplicidade o atributo de uma descrição sintática de entendimento relativamente fácil. Assumindo-se que isso seja uma condição necessária, mas não suficiente, para o aprendizado, descrições sintáticas simples são aquelas que contêm poucos constituintes oracionais, tal que o aprendiz é capaz de acompanhar uma explicação sem perder a coerência dela.

Por isso, árvores que podem cumprir esse requisito são as que se encaixam nos seguintes grupos: (1) aquelas de orações bem formadas, uma vez que elas instanciam um conjunto finito, mas não extenso, de regras de produção [3][8]; (2) as de orações com poucos constituintes, pois essas descrições permitem a visualização do todo; e (3) as descrições anotadas por sistema de rótulos sugestivos, o qual facilita o reconhecimento e interpretação das etiquetas morfossintáticas com quase nenhum aprendizado.

O corpus *Araucárias* incluem árvores com esse perfil e por isso atende à condição da simplicidade. Como é possível notar nas Figuras 4 e 5, as árvores são pouco extensas, as regras visualmente claras e, especificamente, as notações morfossintáticas bem sugestivas. Aliás um cuidado durante a construção desse corpus foi substituir as etiquetas morfossintáticas da GG por etiquetas mais sugestivas, tais como "N" (Nome) por "Sub" (Substantivo); "Det" (Determinante) por "Art" (Artigo); entre outros.

Por essa razão, os dados de anotação não precisam ser necessariamente realistas como as expressões da linguagem cotidiana. Assim, eles podem ser obtidos sinteticamente através das ferramentas de Inteligência Artificial Generativa. Isso não é nenhum demérito para este projeto, já que a natureza do corpus permite o uso dessas ferramentas. Ademais, o uso de inteligência artificial para a construção de corpora reduz os esforços cognitivos e os custos financeiros, além é claro do tempo gasto para o levantamento de dados.

Especificamente, no tocante à redução dos esforços cognitivos e de tempo de busca, trabalhar com a inteligências artificiais permite controlar a produção dos dados (ideais) para a confecção de corpora. Uma vez que um corpus prima por ser representativo, informativo e balanceado [10], isso tudo pode ser conseguido por meio de comandos de prompt adequados, como os que foram empregados no *Araucárias*. A princípio, uma amostra de 6.000 árvores, representando 20 padrões de estrutura sintática, é suficiente para a criação de atividades metalinguísticas e, desde que treinados, também para a elaboração de aplicações web e APIs.

Em vista do exposto, os corpora *Floresta* e *Porttinari* seriam, a princípio, pouco adequados ao ensino formal da sintaxe do português. Como compreendem árvores baseadas em orações reais do português escrito, as quais são complexas, rebuscadas e descritas por meio de sistema de notação universal, que é pouco sugestivo, esses corpora pecam em ser simples para o ensino;

Como dito, eles são assim porquanto foram elaborados para atender às necessidades da indústria e não propriamente da educação. Todavia, isso não significa dizer de forma alguma que eles não possam atender às demandas escolares. Com muitos ajustes e uma grande dosagem de criatividade, essas informações podem ser transformadas em aplicações pedagógicas úteis. Seja como for, em comparação com o *Floresta* e o *Porttinari*, o corpus *Araucárias* é atualmente mais condizente com o ensino.

5.3 **Árvores com descrições evocativas.**

Diferentemente dos caracteres comentados anteriormente, ser intuitivo e ser simples, o caráter evocativo de descrições sintáticas é fácil de ser definido. Uma descrição é evocativa se ela ativa experiências prévias do aprendiz com o aprendizado da sintaxe do português. Se, de um lado, a definição de evocação é clara, de outro, a verificação desse caráter requer esforços científicos adicionais de medição e avaliação. Por isso, esse caráter é tratado neste projeto como uma assunção e não como um fato. Não obstante, há fortes razões para isso.

Em primeiro lugar, as descrições contidas nas árvores do *Araucárias* são *grosso modo* similares às descrições permitidas pela GT. Isso quer dizer que uma pessoa que esteja aprendendo a sintaxe pelo viés desse corpus e que seja capaz de perceber a referida similaridade estaria se lembrando de práticas metalinguísticas prévias, experienciadas nas aulas de português do ensino médio. Isso acontece porque tais árvores são não disruptivas.

Nesse caso, mesmo na situação em que os aprendizes não consigam notar a similaridade, o desenho da árvore é suficientemente isomórfico com as descrições tradicionais que o docente terá mais facilidade em mostrar as semelhanças e em recuperar o curso normal do processo de aprendizagem. E isso aponta para a próxima razão.

Em segundo lugar, as árvores do *Araucárias* são potencialmente ilustrativas das análises sintáticas tradicionais. Devido à isomorfia fraca entre descrições, mesmo que isso não seja imediatamente percebido, tais árvores permitem uma abordagem que revela as referidas similaridades a partir de conhecimentos prévios dos estudantes. O docente, nesses casos, precisa apenas instigar a curiosidade e a percepção dos aprendizes. Esse fato aponta para a próxima e última razão.

Finalmente, em terceiro lugar, percepção e cognição não são processo separados, mas sim imbricados. Com efeito, expectativas prévias, baseadas em conhecimentos armazenados, orientam como os eventos vão ser apreendidos. E, no sentido inverso, novos insights perceptuais podem alterar concepções arraigadas.

A linguagem verbal cumpre um relevante papel nesse processo, como têm mostrado algumas pesquisas [11]. Com efeito, as palavras têm o poder de revelar novos aspectos das experiências e fazer mudar as crenças sobre o mundo. Se se considerarem as memórias como parte desse processo perceptual, então a percepção é em muitos aspectos evocativa também.

Disso deriva que árvores sintáticas capazes de evocar práticas prévias têm mais chances de facilitar a compreensão da estrutura sintática das orações do que aquelas que, ao contrário, são quase completamente arbitrárias. Esse caráter evocativo diz respeito aos procedimentos analíticos, mas especialmente à terminologia empregada. Se é verdade que as palavras têm o poder de facilitar a percepção, e se a terminologia é uma forma de linguagem – isto é, uma língua de especialidade [12] –, então o sistema de notação cumpre esse papel de facilitador da percepção.

Assim sendo, no que diz respeito ao ensino de sintaxe, *Araucárias* leva vantagem em relação a seus congêneres, o *Floresta* e o *Porttinari*. Que isso é o caso pode ser verificado considerando a metalinguagem adotada por esses corpora, a qual é quase totalmente arbitrária e com pouco poder de evocação.

5.4 **A relevância da ciência da linguagem**

Diante do exposto, uma questão capciosa poderia surgir: se as descrições da GT devem ser evocadas por árvores sintáticas, por que não continuar ensinando a GT ao invés de se criar um construto misto, tradicional e científico, para se fazer isso? Ou seja, ensinar sintaxe tradicional não seria cognitivamente mais econômico do que uma abordagem científica que "contraditoriamente" seja evocativa de experiências metalinguísticas tradicionais que ela mesma condena?

A resposta a isso é obviamente uma negação. A ciência existe para depurar as explicações dos fatos, eliminando vieses, falsas crenças, preconceitos etc. Com a ciência linguística não é diferente. A Linguística nasceu com esse objetivo e como uma reação à abordagem tradicional. No domínio da educação, ela apoia um estudo mais objetivo da linguagem, sem as inconsistências que caracterizam a GT.

Nesse sentido, o ensino da sintaxe deve necessariamente se orientar pelo conhecimento linguístico, como o da GG. Porém, isso não significa que a abordagem deva ser arbitrária e disruptiva, o que dificulta a aprendizagem. Essa afirmação soa como uma contradição, mas no fundo não o é. Na realidade, é uma maneira de ensinar ciência baseada em experiências anteriores, mesmo que essas sirvam apenas como ponto inicial.

Araucárias certamente não descuidou desse problema e respondeu a ela. Isso acontece quando ele inclui uma concepção pedagógica que, em última análise, vem de empréstimo do trabalho de Souza e Silva, e Koch (1995). Os ajustes que foram feitos na etiquetagem morfossintática deixaram o corpus ainda mais sugestivo e apropriado para o estudo em sala (Figura 4).

Contudo, uma ponderação deve ser considerada: o fato de o *Araucárias* ser condizente com o ensino não significa que ele seja melhor que seus correlatos. Na verdade, se o propósito for embasar

aplicações para a indústria, atualmente seus concorrentes se sobressaem a ele.

6 **Considerações finais**

Como foi visto, o ensino de sintaxe do português brasileiro, usando recursos de PLN, como os corpora de árvores sintáticas, requer amostras que, desde seu nascimento, devem incluir uma concepção pedagógica. O corpus *Araucárias*, inspirado em trabalhos de linguística aplicados ao português [5], pode cumprir bem essa função.

Trata-se de um corpus que se pretende didático, logo desprovido da sofisticação e complexidade que aqueles voltados para a indústria apresentam. Como se viu, ele é avaliado por sua capacidade de satisfazer três condições fundamentais para a instrumentalização do conhecimento do vernáculo: ser intuitivo, ser simples e ser evocativo.

Com efeito, ele satisfaz com folga essas condições, porquanto seus dados são produzidos em condições muito bem controladas. Ou seja, são dados sintéticos produzidos via inteligência artificial. Porém, não há nenhum demérito nisso porque o ensino deve partir do mais simples para o mais complexo, do mais inteligível para o mais intrincado, e do mais intuitivo para o mais abstrato.

Seja como for, ele pode funcionar em diferentes aplicações de ensino. Algumas delas podem ser: recurso para a produção de material didático; banco de dados para APIs de análise sintática; e *dataset* para aprendizagem de máquina. Nesse último caso, os resultados podem ser usados em aplicativos voltados para a educação.

REFERÊNCIAS

[1] Magali S. Duran, Lucelene Lopes, Maria das Graças V. Nunes e Thiago A. S. Pardo. 2023. The Dawn of the Porttinari Multigenre Treebank: Introduction its Journalistic Portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL'23)*, SBC, Belo Horizonte, p.115-124. DOI:https://doi.org/10.5753/stil.2023.233975

[2] Cláudia Fritas, Paulo Rocha e Eckhard Bick. 2008. Floresta Sintá(c)tica: Bigger, Thicker, Easier. In: António Teixeira et al. (Eds.). 2008. In *Proceedings of the 8th International Conference on Computational Processsing of the Portuguese Language*. Spring-Velarg Berlin Heidelberg, 216-219. DOI: https://doi.org/10.1007/978-3-540-85980-2_23

[3] Noam ChomsKy. 1957. *Syntactic Structures* (2ª ed.). Berlin: Mouton de Gruyter, 2002.

[4] Evanildo Bechara. *Moderna Gramática Portuguesa*. Ed. Revista e Ampliada. Rio de Janeiro: Lucerna, 2002.

[5] Rocha Lima. *Gramática Normativa da Língua Portuguesa*. 41 ed. Rio de Janeiro: José Olympio, 2001.

[6] Winfried Busse e Mário Vilela. 1986. *Gramática de Valências*. Coimbra: Almedina.

[7] Gisely G. de Castro. Teoria Gerativa: Contexto Histórico, Desenvolvimentos Recentes e Compromissos Futuros. *Caderno CESPUC*, n. 33, p. 21-35, 2018. Disponível em: https:// d1wqtxts1xzle7.cloudfront.net/108533437/14115-libre.pdf? Acesso em: 24 nov. 2024.

[8] Maria C. P. de Souza e Silva, Ingedore G. V. Koch. 1995. *Linguística Aplicada ao Português: Sintaxe* (6ª ed.). São Paulo: Cortez.

[9] spaCy. 2024. *Industrial-strength Natural Language Processing in Python.* Disponível em: https://spacy.io/models. Acesso em: 24 nov. 2024.

[10] James Pustejovsky e Amber Stubbs. 2013. *Natural Language Annotation for Machine Larning*. CA, USA: O'Reilly Media, Inc.

[11] Lera Boroditsky. How language shapes thought. *Scientific American*, v. 304, n. 2, p. 62-65, 2011.

[12] Isabel T. M. Gil. Algunas considerações sobre línguas de especialidade e seus processos lexicogénicos. *Máthesis*, n. 12, p. 113-130, 2003. DOI: https://doi.org/10.34632/mathesis.2003.n12