Projeto de Pesquisa do Programa de Iniciação Científica ou Tecnológica

Banco de dados para análise sintática em sentenças do português brasileiro

Willian Emerson Afonso Pacheco

Orientador: **Manoel Francisco Guaranha**

São Paulo

Abril, 2021

1. Resumo do Projeto

A Teoria Gerativista de Chomsky teve profunda influência sobre a linguística moderna e também sobre o interesse pela automatização da linguagem humana. Como consequência desta, surgiu um campo de estudos voltado ao processamento de linguagem natural (PLN). Estruturas de dados para armazenar o léxico presente em dicionários humanos precisaram ser elaborados e dentre os autômatos desenvolvidos para português brasileiro, o PALAVRAS de Erick Bick e os analisadores sintáticos do Núcleo Interinstitucional de Linguística Computacional (NILC) tiveram destaque. Tendo por base o Unitex-PB do NILC, o presente trabalho visou construir um modelo de banco de dados léxico e de estruturas sintáticas em português brasileiro para apoiar programas de análise sintática. O estudo foi dividido em três etapas: um estudo referencial sobre as principais estruturas envolvidas na formulação de um *parser*, a modelagem da estrutura entidade-relacionamento (ER) do banco de dados e a implementação através do gerenciador de banco de dados PostgreSQL administrado pelo Dbeaver. Ao final obteve-se um modelo ER próprio para a utilização em compiladores e que abrange catorze categorias sintáticas: substantivos, adjetivos, artigos, preposições, conjunções, numerais, pronomes, verbos, advérbios, prefixos, contrações, siglas, abreviaturas ou interjeições.

2. Introdução

2.1 Histórico e uso da programação para a linguagens naturais

A linguagem humana é um sistema composto por unidades menores que dão sentido e forma a comunicação. Quando falamos ou escrevemos utilizamos uma estrutura própria da língua para expressar ideias e nos comunicarmos. Essa estrutura permite que falantes de uma mesma língua possam se entender utilizando-a como uma base comum para relacionar os diferentes conceitos que podem estar envolvidos na sentença.

A estrutura de todas as línguas pode ser entendida como formada por elementos comuns com funções iguais. Essa é a ideia que fundamenta o trabalho de Noam Chomsky (2015) pai da Teoria Gerativista. Nesta teoria, o autor propõe o que ele chama de Gramática, em letra maiúscula para se diferenciar da gramática normativa e cujo foco é a sintaxe, ou construção de sentenças.

Seu estudo parte da seguinte pergunta: como explicar as frases realizadas e que podem ser realizadas em uma língua? A partir de um intenso estudo de várias estruturas das línguas do mundo ele propõe que por meio do raciocínio dedutivo sobre regras abstratas finitas o falando é capaz de gerar inúmeras sentenças. As menores unidades de sentido de uma sentença são agrupadas dentro de Sintagmas.

A Teoria Gerativista de Chomsky teve profunda influência sobre a linguística moderna e também sobre o interesse pela automatização da linguagem humana. Ao fornecer uma estrutura de regras simples capaz de formar sentenças complexas ela tornou possível que linguagens próprias pudessem ser construídas. As linguagens humanas (naturais como são chamadas) receberam então um novo corpo em linguagens programadas ou de programação.

Ao longo das décadas que se passaram, a Teoria Gerativista foi gradualmente incrementada e as linguagens de programação cresceram e se desenvolveram no meio acadêmico e técnico. Um campo de processamento de linguagem natural (PLN) surgiu e estruturas de dados para armazenar o léxico presente em dicionários humanos precisaram ser elaborados.

Muitas experiências surgiram a partir desse campo. Buscou-se criar desde tradutores a autômatos que pudessem se engajar em conversas com pessoas reais. No centro dessas ferramentas estão os analisadores sintáticos (ou *parsers,* em inglês), autômatos responsáveis por descobrir os sintagmas presentes em sentenças fornecidas por um usuário ou outro programa.

Até o início dos anos 80 a preocupação com o desenvolvimento de léxicos e bases de dados lexicais era baixa, o que tornava a reutilização algo bem dificil (MUNIZ, 2004). A língua não é estática e depende do contexto e dos indivíduos que a utilizam. Uma língua pode conter mais de 1 milhão de unidades léxicas (entre palavras, locuções e morfemas) e esse número cresce constantemente a medida que há o intercambio e criação de léxicos pelos falantes que a utilizam. Uma estrutura comum é, portanto, uma consequência natural do desenvolvimento de estudos no campo das PLN.

2.2. A estrutura do DELAF-PB

O Delaf-PB é formado por 9 milhões de linhas dentre palavras simples e flexionadas. Cada linha é uma entrada de texto terminada com um caractere de quebra de linha (\n) e formada por até cinco áreas de acordo com o padrão abaixo:

palavra,canônica.Classe+traços:flexão

- Uma vírgula (“,”) é utilizada para separar a *palavra* em sua forma flexionada da sua forma *canônica.* Para substantivos e adjetivos a forma canônica está geralmente na forma masculina; para verbos, está em sua forma infinitiva;

- Um ponto (“.”) representa uma classe gramatical. Um verbete pode pertencer a categorias dos substantivos, adjetivos, artigos, preposições, conjunções, numerais, pronomes, verbos, advérbios, prefixos, siglas, abreviaturas ou interjeições. Cada verbete pode estar classificado em mais de uma classe gramatical, neste caso há uma entrada para cada classe;
- Um sinal de adição (“+”) representa a sequência de uma classe gramatical e uma informação semântica ou pronome átono. Uma classe pode ser associada aos seguintes conteúdos: próprio(a), indefinido(a), coletivo(a), demonstrativo(o), relativo(a), interrogativo(a), tratamento(a), possessivo(a) e pessoal.
- Por fim, o sinal de dois pontos (“:”) é utilizado para delimitar informações de flexão, tempo, forma, valor, gênero, número, grau, pessoa. Um verbete pode conter muitas flexões.

3. Objetivos Gerais e específicos

Objetivo geral:

Construir um modelo de banco de dados de léxicos e estruturas sintáticas em português brasileiro para apoiar programas de análise sintática;

Objetivos específicos:

Elaborar uma proposta de estruturas sintáticas para o português brasileiro.

Elaborar um modelo entidade-relacionamento de estruturas sintáticas para o português brasileiro.

4. Revisão Bibliográfica

A produção científica sobre a automatização do processo de análise da estrutura sintática e semântica da língua portuguesa passou por um período de intenso crescimento no início do século XX. Muitos destes estudos foram desenvolvidos entre a década de 1990 e 2008 (BICK, 2000; GREGHI, 2002; MUNIZ, 2004; DIAS-DA-SILVA, 2008, entre outros).

Grande parte dos *resultados* para português brasileiro encontra-se reunido no projeto AC/DC (Acesso a corpos/Disponibilização de corpos) da Linguateca, um centro de recursos para o processamento computacional da língua portuguesa. Essa iniciativa foi desenvolvida para centralizar os recursos disponíveis e facilitar a comparação entre eles. Todos os resultados das pesquisas dentro do escopo do projeto foram disponibilizadas no site do projeto junto a uma interface gráfica simples que auxilia na consulta de cada um.

Os *corpora* reunidos no AC/DC tiveram sua construção baseada em *corpus* anotados automaticamente pelo *parser* PALAVRAS, do pesquisador Eckhard Bick. O PALAVRAS é , como define Bick (2000), um analisador sintático e semântico que utiliza uma notação própria baseada no paradigma metodológico da *Constraint Grammar* (CG). A CG utiliza a dependência de contexto entre as palavras numa sentença para gerar regras gramaticais. No PALAVRAS cada uma das regras age como uma estrutura condicional com a função de adicionar, remover, selecionar ou substituir uma etiqueta ou um conjunto de etiquetas para um determinado contexto. Essa estratégia reducionista permitiu ao *parser* restringir-se a inicialmente 8 mil regras gramaticais possíveis, comparativamente menos do que outras metodologias.

Quando somadas, são mais de 1 bilhão de palavras etiquetadas entre as mais diferentes versões do *parser* que utilizaram muitos tipos de fontes de dados para serem anotados (jornais, textos literários, mensagens de correio, entre outras). O trabalho de Bick causou por este motivo uma grande influência sobre a padronização, a elaboração e a organização dos dados utilizada nos *corpora* de palavras em português brasileiro.

Como um trabalho decorrente do PALAVRAS, o Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo desenvolveu, desde sua criação em 1991, vários aplicativos e recursos linguísticos para o português brasileiro. Entre os aplicativos desenvolvidos, destaca-se o Diadorim que incorpora informações presentes no léxico do NILC a um banco de dados relacional que centraliza todas as informações lexicais; e o Unitex-PB que utiliza um dicionário denominado DELA onde cada entrada é descrita em uma estrutura padrão.

A construção de estruturas de armazenamento de dados léxicos tem tido pouca atenção em tempos recentes. Muitos dos estudos atuais têm se focado em métodos de processamento mais eficientes e modernos que utilizam inteligência artificial com modelos matemáticos em vez de regras gramaticais estruturadas. De forma geral tem sido um consenso que estruturas de armazenamento de dados boas devam permitir que sejam modificadas e reutilizadas por um público variado que vai linguísticas a programadores e matemáticos.

Por este motivo se consolidaram duas formas de bases: dicionários tratáveis por máquinas e bases de dados lexicais. Os dicionários contêm descrições do conhecimento lexical em linguagem simbólica formal que são traduzidos por um sistema específico e permitem conteúdos que normalmente estão implícitos em dicionários para uso humano. O Dicionário DELA do NILC é um exemplo deste tipo de armazenamento. Eles têm sido cada vez menos utilizados porque dependem de programas específicos para interpretação.

As bases lexicais, no entanto, têm crescido por serem desenvolvidas em formato de banco relacional e em linguagem SQL, o que facilita a padronização das consultas. São arquivos de computador compostos basicamente por registros organizados logicamente em uma estrutura que permite fazer consultas e alterar os dados presentes. Sua estrutura usualmente contém um

conjunto de tabelas, dentre as quais uma contém lexemas com alguma informação diretamente relacionada a eles e outras se relacionam a ela.

5. Metodologia

Pretendo inicialmente compreender o estado da arte dos estudos sobre as classes gramaticais tendo em Othero (2009) e em um estudo dirigido com o Professor Doutor Manuel Francisco Guaranha uma referência inicial. Essa etapa servirá para embasar os principais conceitos envolvidos atualmente na formulação de um *parser* sintático.

Utilizarei o dicionário Unitex-PB, recurso construído com base no sistema francês DELA. O dicionário está disponível na internet em formato de texto e tem a documentação bem detalhada no trabalho de Muniz (2004). Ele servirá como uma base de dados acessível e como um ponto de partida para a estrutura gramatical. O objetivo desta etapa é relacionar as classes gramaticais apreendidas durante a etapa de estudo do estado da arte.

A escolha da estrutura sintática adotada pelo *parser* é um momento bastante importante da construção porque ela é quem apoia a estrutura do banco. Chomsky em seu estudo inicial já apontava a formação inerentemente recursiva e arbórea das partes que compõe a estrutura da língua. É uma estrutura relacional por definição, cujas entradas são pouco atualizadas ao longo do tempo e com elementos (sintagmas) que tem graus de dominância e subordinação variados. Um banco de dados relacional é a estrutura de armazenamento mais adequada para essa tarefa, e será, portanto, adotado.

Definidas as classes gramaticais e seus elementos correlacionados (gênero, número, grau, etc) seguirei para a construção do modelo relacional. Aqui, serão definidas as entidades-relacionamentos (ER) de acordo com boas práticas de normalização dos dados dadas pelas com as Formas Normais de Codd (1972 apud LAKE and CROWTHER, 2013) e pela ampliação destas por Bryce-Codd. Muitos modelos ER podem representar uma mesma estrutura, por isso o desafio é ter um produto acessível a linguistas e programadores interessados na tarefa de análise sintática. Esses públicos têm conceitos muito próprios de suas áreas e serão considerados durante a escolha do modelo de dados.

A implementação, por fim, será feita com o sistema gerenciador de banco de dados objeto-relacional (SGBD) PostgreSQL e apoiada pelo Dbeaver, uma ferramenta para administração rápida e de código aberto. Todos os algoritmos que forem necessários para processar o dicionário e transformar as ocorrências do Unitex-PB em entradas do modelo de dados escolhido serão construídos em linguagem Go.

6. Resultados

6.1 Modelagem do banco de dados

O processo de construção do banco foi orientado pelas boas práticas de normalização dos dados de acordo com as Formas Normais de Codd (1972). Como bem descreve LAKE e CROWTHER (2013) o principal objetivo da normalização é aumentar a performance e diminuir a necessidade de armazenamento através da diminuição de redundâncias de dados. Para chegar aos resultados finais, porém, foi necessário passar por diversas experimentações que envolviam não só boas práticas, mas também um olhar para a otimização da busca.

Partiu-se de uma grande estrutura monolítica que contemplasse todas as palavras e subcategorias existentes no DELAF-PB. Tal qual um dicionário físico, isso trouxe o problema da repetição excessiva de uma mesma palavra mesmo que um só atributo fosse diferente. Isso foi um entrave especialmente para tempo e traço semântico em verbos, para os quais pode haver muitos tempos e muitos traços como se vê no exemplo na Tabela 1.

Tabela 1 - Exemplo de uma estrutura monolítica para a palavra construir

Lexema	Canônica	Classe	Tempo	Pessoa	Número
construir	construir	V	U	1	s
construir	construir	V	U	3	s
construir	construir	V	W	-	-
construir	construir	V	W	3	s

Essa seria uma clara violação da Primeira Forma Normal apesar de simplificar bastante a construção de consultas ao banco. Foi necessário, portanto, realizar um mapeamento das categorias e subcategorias para verificar a cardinalidade das relações entre o Léxico e todas as categorias gramaticais.

Partiu-se da premissa que o Léxico poderia ser uma tabela central para estruturar as relações. Um léxico é por definição composto de uma palavra canônica, forma-base da entrada no dicionário e por um lexema, forma flexionada que representa a unidade léxica que contém as marcas gramaticais (BIDERMAN, 1984). Esses dois componentes são atributos obrigatórios para a existência de uma entrada no dicionário e uma unidade lexical se diferencia da outra também por esses dois componentes.

Realizando uma proposta para o problema levantado na Tabela 1, a palavra *construir* poderia ser resumida em uma única entrada. Pragmaticamente falando, a criação de uma chave-

primária com dois valores do tipo *string* ou *char* ainda que obedeçam a boas práticas poderia trazer mais riscos à implementação. Do ponto de vista da compilação de código, caracteres de acentuação podem gerar ainda problemas de incompatibilidade entre sistemas e padrões de codificação (por exemplo, em um documento não propriamente codificado em formato *utf-8*).

Existiu aqui também um problema muito mais teórico do que prático: o conjunto lexima+forma-canonica pode se repetir porque podem ter classificações diferentes em contextos diferentes. O lexema *casa* por exemplo pode ser classificado como verbo ou substantivo. Além disso, lexema e forma-canônica tem cardinalidade NxN o que demandaria uma tabela associativa e a criação de duas tabelas novas, o que na realidade não eliminaria as duplicações porque ambas podem ser iguais (como em *casar* - verbo infinitivo - e *casar* forma-canônica). A solução lógica foi a separação de todas as entradas em uma tabela própria (Tabela *Palavra*), a criação de uma entidade associativa com o lexema e a forma-canônica como chaves estrangeiras (Tabela *palavra_lexico*) e a conversão do léxico em uma tabela auxiliar (tabela *Lexico*) que recebe o atributo *unidade_lexical* interligando-se ao restante (Imagem 1).

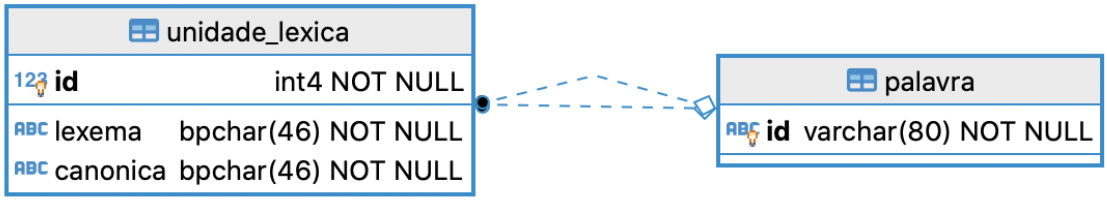


Imagem 1 - Relacionamento entre Palavra e Unidade Lexical. Uma unidade léxica é formada por um lexema e uma forma-canônica.

Isso faz bastante sentido porque uma das características da língua é a flexibilidade na formação do léxico. As pessoas recombinam as palavras para dar origem a novas unidades lexicais a partir de um repertório que cresce e muda constantemente. Não há nenhuma garantia que o lexema *casa* não possa ser empregue no futuro com sentido outro além de verbo ou substantivo porque isso não importa ao falante. Quem se interessa pela gramática são os linguistas e alguns programadores. O restante irá recombinar esses blocos livremente e isso poderia impactar profundamente na performance de um banco de dados já que essas mudanças e inserções de novas palavras acontece constantemente. Então o Léxico (esse conjunto de unidades e categorias gramaticais) e o repertório precisam ser independentes, o que corrobora a separação em tabelas diferentes. Diminui-se assim o problema prático porque a busca pode ser realizada unicamente com a palavra (e não mais com o conjunto lexema+forma-canonica) e soluciona-se o problema das ambiguidades (ao menos no escopo do banco) porque diminui as duplicações entre lexema e forma-canônica pois ambos estão reunidos na tabela *Palavra*.

O DELAF traz ao todo oito grupos gramaticais: classe, gênero, número, grau, tipo, forma, pessoa e tempo. No caso de algumas das categorias (pessoa, gênero, grau e número) a repetição deriva-se de tais atributos terem dependências multivaloradas. Em alguns casos, duas classes podem se relacionar com um único grupo, como é o caso dos substantivos e adjetivos que compartilham o atributo gênero. Essas relações cruzadas formam na prática uma rede de conexões semelhantes a um *grapho* direcionado que pode ser representado por uma matriz de adjacências (Tabela 2).

Tabela 2 - Matriz de adjacências para Traço Semântico

Categoria	Descrição	Substantivo	Adjetivo	Artigo	Numeral	Pronome	Sigla	Abrev	Verbo
Gênero	Masculino	1	1	1	1	1	1	1	0
	Feminino	1	1	1	1	1	1	1	0
Número	Plural	1	1	1	1	1	1	1	0
	Singular	1	1	1	1	1	1	1	0
Grau	Aumentativo	1	1	0	0	0	0	0	0
	Diminutivo	1	1	0	0	0	0	0	0
	Superlativo	1	1	0	0	0	0	0	0
Pessoa	Primeira pessoa	0	0	0	0	1	0	0	1
	Segunda pessoa	0	0	0	0	1	0	0	1
	Terceira pessoa	0	0	0	0	1	0	0	1

O número 1 (um) representa uma ligação e o número 0 (zero) a ausência desta.

A presença de uma adjacência é representada na tabela pela sequência de números 1 (um) nas células no eixo horizontal. Na teoria de grafos isso representa uma ancestralidade em comum, tal qual dois filhos de um mesmo pai e por isso podem ser generalizados em um grupo. Isso dialoga com a proposta de Bryce-Codd (1972 apud LAKE and CROWTHER, 2013) e sua Quarta Forma Normal, que diz que quando três ou mais atributos são parte de uma chave composta pode ser necessário decompor a tabela em duas outras. Nesse caso, a linguística já reúne essas categorias dentro do grupo definido como *traço semântico.* Porém, conceitualmente um traço é

uma característica da palavra dentro de um contexto, como no exemplo acima da unidade *casa*. O traço só existe no contexto de uma classe porque é a classificação da palavra que fornece seu significado semântico, já que uma palavra pode ao mesmo tempo estar em sua forma-canônica, representar um lexema ou pertencer a muitas classes. Dessa forma, uma unidade lexical ao ser associada a uma classe permite que um traço possa existir. Dada a cardinalidade entre classe e traço ser NxN optou-se por criar a tabela intermediária e assim evitar as duplicações observadas na matriz de adjacências (Imagem 2).

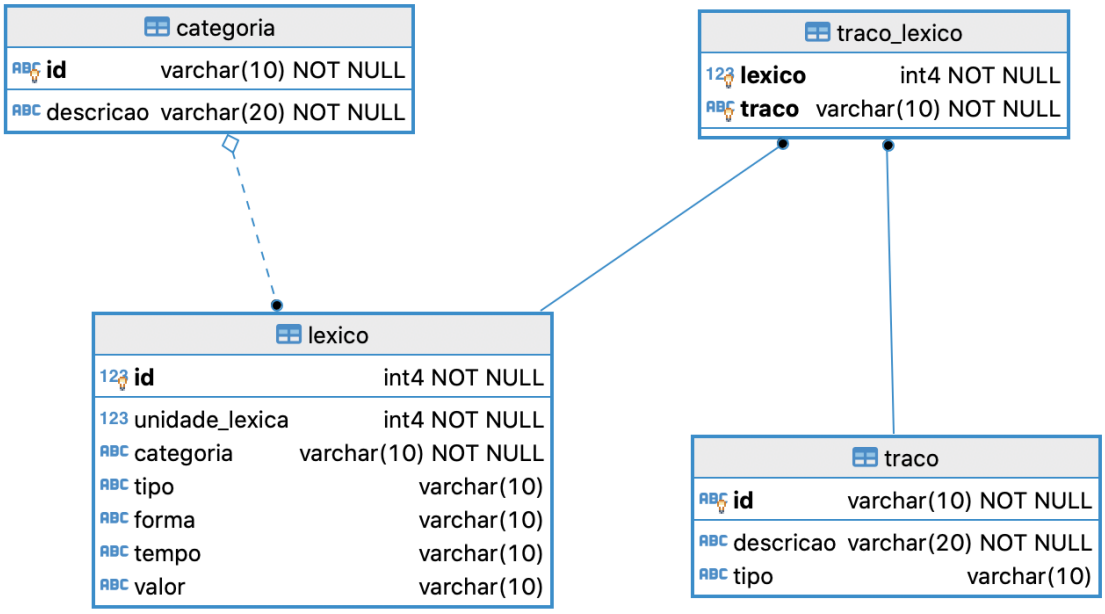


Imagem 2 - Relacionamentos entre traços e classes gramaticais. A classificação de uma palavra em uma classe (atributo obrigatório) permite que um ou mais traços semânticos possam ser associados a ela.

No restante das categorias a complexidade de relacionamentos é muito menor. Ao todo, forma, tipo e tempo têm 84 valores possíveis sem adjacências entre as classes gramaticais. Optou-se portanto por manter essas denominações e criar tabelas homônimas com as letras que simbolizam a categoria como identificador e uma descrição por extenso como *placeholder*. A imagem 3 demonstra o padrão através da tabela *Forma*.

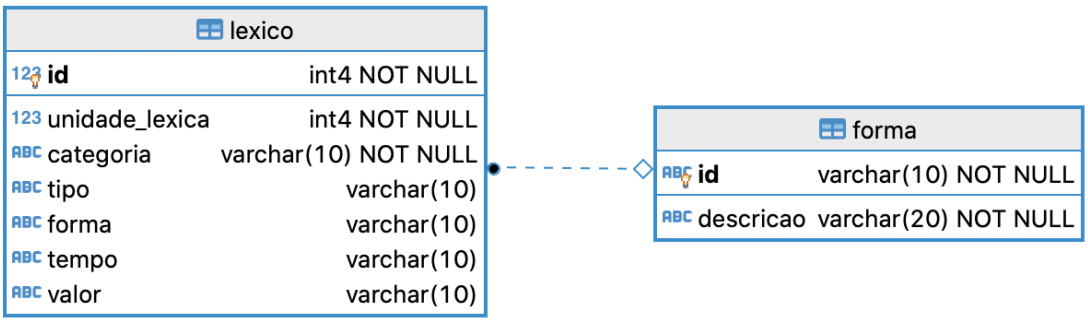


Imagem 3 - Exemplo de padronização para o restante das categorias gramaticais. Cada categoria contém um símbolo (chave primária) e uma descrição; tipo, forma, tempo e valor não são atributos obrigatórios nesse modelo.

Um último entrave encontrado foi a existência de combinações e contrações de palavras, representadas no DELAF pela união de duas classes pela letra *X* (como em *PREPXART* para o lexema *da*). Nem combinações e nem contrações envolvem diretamente categorias gramaticais e por definição tem independência em relação a elas. Elas podem ser definidas genericamente como a junção de uma preposição com palavras de outras classes gramaticais que originam um novo vocábulo. Sem perda de seu conceito original foi possível simplesmente agrupá-las em uma nova entidade com cada entrada identificada pelo prefixo e sufixo correspondente.

O modelo Entidade-Relacionamento completo encontra-se no anexo 1.

6.2. População do banco de dados

Para popular o banco foram necessárias algumas alterações do dicionário original para padronizar a simbologia e facilitar o processamento. Uma primeira dificuldade, foi a opção de (MUNIZ, 2004) por manter o traço semântico e pronomes associados ao mesmo símbolo "+". Na lógica que optamos, e conforme descrito no tópico anterior, as classes gramaticais são organizadas em entidades separadas das demais. Na descrição original do UNITEX o sinal equivalente também se restringe a traços semânticos.

O DELAF apresenta as formas flexionadas dos verbos na forma da expressão regular .**V\+PRO.** em que um verbo se justapõe a um pronome átono. As linhas que pertenciam a este padrão foram retiradas do dicionário original e não foram utilizadas na população do banco.

As entradas classificadas como numerais ou artigos também foram descritas com o sinal "+" pela expressão regular .**DET\+Art.** e .**DET\+Num.**, em que Art representa o artigo, Num o numera e DET os determinantes. Esses símbolos são representações de classes e não de traços semânticos. Seguindo a lógica de padronização, todas as linhas que continham estas informações

foram substituídas pelos símbolos ART e NUM (em caixa alta). Ainda que haja uma grande discussão sobre a utilização dos determinantes para englobar essas classes, esta escolha não afeta a estrutura proposta no tópico 6.1 pois a separação do traço e classe como diferentes entidades não impede que no futuro outros olhares possam ser utilizados para popular o banco.

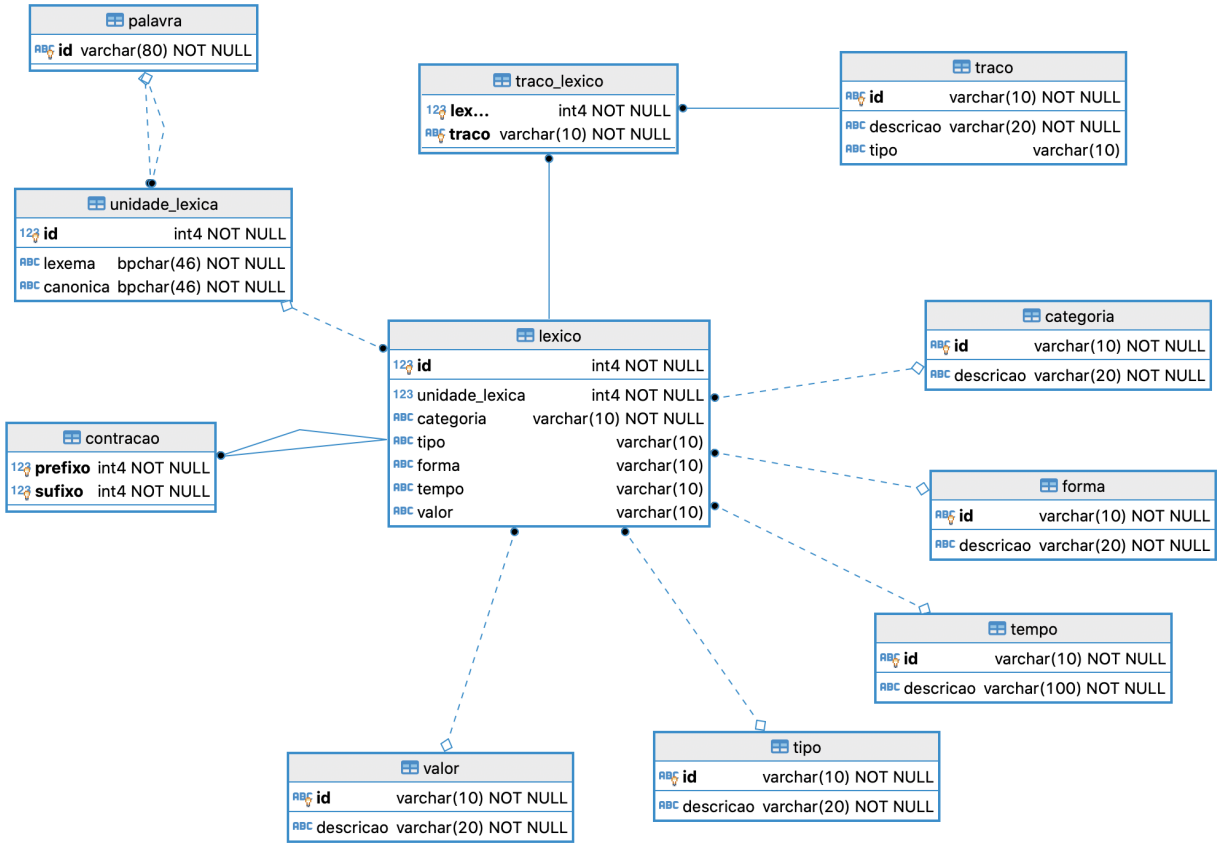
Dentro dos traços semânticos restantes optou-se ainda por inserir categorias com valor denotativo para numerais: cardinais, ordinais, multiplicativos e fracionários. Originalmente essas categorias foram organizadas junto às flexões. Em um olhar mais afundo, contudo, observou-se que o traço semântico foi utilizado como categoria ampla que diz respeito a tipos gramaticais e flexões (MUNIZ, 2004). Os símbolos utilizados após o sinal "+" referiam-se somente a tipos gramaticais (e não a traços semânticos de forma geral) e aqueles utilizados após o sinal ":" englobaram todas as categorias restantes que estão associadas às flexões. Isso levou a classificar os tipos dos numerais como categorias associadas ao primeiro símbolo.

Na prática essa definição teve implicações na construção do algoritmo porque dentro do dicionário uma palavra aparecia com um dos tipos listados e nunca mais do que isso, o que demonstra que pertenciam ao mesmo grupo e uma condicionalidade estaria envolvida no processamento. Os tipos resultantes foram reunidos no Anexo 2.

Por fim, restou-se apenas dividir as flexões restantes em suas categorias gramaticais correspondentes Anexos 3 a 5. Seguindo a lógica expressa no tópico anterior, o gênero, número, grau e pessoa pertenceram uma entidade genérica de Flexão.

7. Anexos

Anexo 1 - Modelo Entidade-Relacionamento do banco de dados



Anexo 2 - Tipos gramaticais utilizados no banco

Tipo	Símbolo

cardinal	c
ordinal	o
multiplicativo	m
fracionario	X
definido	Def
indefinido	Inde
coletivo	Col
demonstrativo	Dem
relativo	Rel
interrogativo	Int
tratamento	Tra
possessivo	Pos
pessoal	Pes
coordenativa	Cd
subordinativa	Sub
correlativa	Cor
proprio	Pr

Anexo 3 - Flexões gramaticais utilizadas no banco

Flexão	**Símbolo**
masculino	m

feminino	f
plural	p
singular	s
aumentativo	A
diminutivo	D
superlativo	S
1	primeira pessoa
2	segunda pessoa
3	terceira pessoa

Anexo 4 - Formas gramaticais utilizadas no banco

Forma	**Símbolo**
acusativa	a
dativa	d
nominativa	N
obliqua	O
reflexivo	R

Anexo 5 - Tempos verbais utilizados no banco

Tempo	**Símbolo**
infinitivo	W

gerundio	G
participio	K
presente do indicativo	P
pretérito imperfeito do indicativo	I
pretérito perfeito do indicativo	J
futuro do presente do indicativo	F
preterito mais que perfeito do indicativo	Q
presente do subjuntivo	B
imperfeito do subjuntivo	T
futuro do subjuntivo	U
futuro do pretérito	C
imperativo	Y

8. Referências Bibliográficas

MUNIZ, M. C. M. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB*. Dissertação de Mestrado. Instituto de Ciências Matemáticas de São Carlos, USP. 72p. 2004. Disponivel em: http://ladl.univ-mlv.fr/brasil/bibliografia/oto/DissMuniz2004.pdf

BICK, Eckhard. The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, Aarhus University Press, 2000. Disponivel: https://visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf

DIAS-DA-SILVA BENTO CARLOS; FI FELIPPO, Ariani; Volpe Nunes, Maria das Gracas. The Automatic Mapping of Princeton WordNet Lexical-Conceptual Relations onto the Brazilian Portuguese WordNet Database. Sixth International Conference On Language Resources And Evaluation, Lrec 2008. Paris: European Language Resources Assoc-elra, p. 1535-1541, 2008. Disponível em: <http://hdl.handle.net/11449/117718>.

GREGUI, J. G.; MARTIN, R. T.; NUNES, M. G. V. Diadorim: a Lexical database for Brazilian Portuguese In. *International Cconference on Language Resources and Evaluation LREC 2002, Las Palmas de Gran Canaria Proceedings of the Third International Conference on language Resources and Evaluation*, Manuel G. Rodríguez and Carmem P. S. Araujo (Eds.), 2002, v. iV, n. , p. 1346-1350. Disponivel em:http://www.nilc.icmc.usp.br/nilc/download/GreghiMartinsNunes.pdf

CHOMSKY, Noam. Estruturas sintáticas. São Paulo: Vozes, 2015.
_____. A Ciência da linguagem. São Paulo: Editora UNESP, 2014.

Othero, Gabriel de Ávila A gramática da frase em português [recurso eletrônico] : algumas reflexões para a formalização da estrutura frasal em português / Gabriel de Ávila Othero. – Dados eletrônicos. – Porto Alegre : EDIPUCRS, 2009. 160 p.

BIDERMAN, M. T. C. Glossário. Alfa, São Paulo, 28 (supl.), 135-144, 1984. Disponível em: < http://seer.fclar.unesp.br/alfa/issue/view/284/showToc>. Acesso em 25 Mar 2021.

LAKE, P.; CROWTHER P. Concise Guide to Databases. Springer. 307p. London, 2013